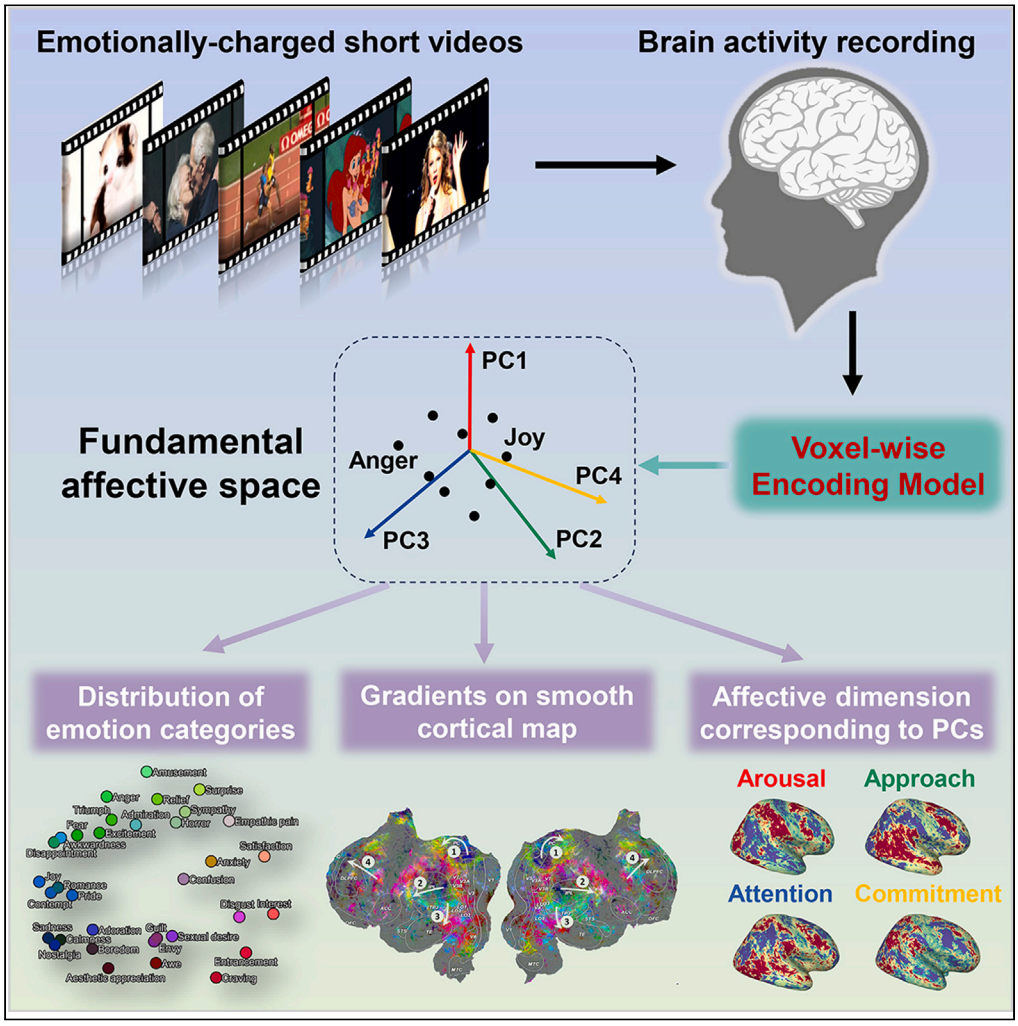
iScience

CellPress
OPEN ACCESS

Article

Topographic representation of visually evoked emotional experiences in the human cerebral cortex



Changde Du,
Kaicheng Fu,
Bincheng Wen,
Huiguang He

huiguang.he@ia.ac.cn

Highlights

Fundamental affective space was constructed from voxel-wise encoding model

Many hypothesized affective dimensions are captured by fundamental affective space

Found that several affective gradients are located across the default mode network

The debate between two views of emotion neural representation is reconciled

Du et al., iScience *26*, 107571
September 15, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.107571

iScience

CellPress
OPEN ACCESS

Article

Topographic representation of visually evoked emotional experiences in the human cerebral cortex

Changde Du,[1,4] Kaicheng Fu,[1,3,4] Bincheng Wen,[2,3] and Huiguang He[1,3,5,*]

SUMMARY

Affective neuroscience seeks to uncover the neural underpinnings of emotions that humans experience. However, it remains unclear whether an affective space underlies the discrete emotion categories in the human brain, and how it relates to the hypothesized affective dimensions. To address this question, we developed a voxel-wise encoding model to investigate the cortical organization of human emotions. Results revealed that the distributed emotion representations are constructed through a fundamental affective space. We further compared each dimension of this space to 14 hypothesized affective dimensions, and found that many affective dimensions are captured by the fundamental affective space. Our results suggest that emotional experiences are represented by broadly spatial overlapping cortical patterns and form smooth gradients across large areas of the cortex. This finding reveals the specific structure of the affective space and its relationship to hypothesized affective dimensions, while highlighting the distributed nature of emotional representations in the cortex.

INTRODUCTION

In our daily lives, we can easily experience a rich variety of emotions, spanning both positive and negative ones. These experiences rely on the intricate encoding and representation of emotional stimuli within the brain. At the core of this cognitive process lies the human emotion system.[1–3]

There are currently two prominent theories regarding the neural representation of emotional experiences: the "locationism" and "constructionism" models. The former proposes that emotional experiences are composed of a finite set of discrete basic emotions (e.g., happiness, anger, fear, and surprise), each of which is encoded independently by highly specialized brain regions.[4] This viewpoint is supported by evidence that focal brain damage leads to specific emotional deficits, particularly concerning the amygdala and fear.[5–7] In contrast, the constructionism model posits that emotion experiences are the result of a series of mental events formed by the interaction of domain-general and large-scale brain networks. There existed some connectivity-based studies[8,9] which explored the dynamic functional connectivity between these networks when people received naturalistic stimuli. A previous study[10] has also shown that these brain networks are not related to a specific emotion category, but to the intensity of emotion experiences. Apart from this, investigating the semantic space in which emotions are represented in the brain can also be deemed as supporting the constructionism hypothesis, in which continuous affective dimensions, such as valence and arousal,[11–13] are involved. This hypothesis is supported by the observation that focusing on affective dimensions may facilitate the interpretation of neuroimaging data.[14,15]

The debate between these two perspectives is a contentious and ongoing issue in affective neuroscience. Multi-voxel pattern analyses (MVPA) of large-scale neuroimaging data have failed to provide a conclusive answer, as they support both discrete emotion category representation[4,16] and distributed networks encoding continuous affective dimensions.[14,15]

Considering the computational constraints of the human brain, it is unlikely that all features of emotional experience are processed by distinct brain regions or networks. A more biologically plausible hypothesis is the encoding of emotional experiences in a gradient-like manner.[17] Evidence supporting this hypothesis would lend more weight to the constructionism model, as the spatial arrangement of distinct gradients would enable the brain to efficiently map various emotional states within a limited range of brain regions. Although neural substrates representing discrete emotion categories or continuous affective dimensions have been identified,[17–19] it remains unclear whether a fundamental affective space underpins the discrete emotion categories in the human brain, and how it relates to the hypothesized affective dimensions.

[1]Laboratory of Brain Atlas and Brain-Inspired Intelligence, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Science, Beijing 100190, China
[2]Center for Excellence in Brain Science and Intelligence Technology, Key Laboratory of Primate Neurobiology, Institute of Neuroscience, Chinese Academy of Sciences, Shanghai 200031, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[4]These authors contributed equally
[5]Lead contact
*Correspondence: huiguang.he@ia.ac.cn
https://doi.org/10.1016/j.isci.2023.107571

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In this study, we developed a novel voxel-wise encoding model[20] to analyze human functional magnetic resonance imaging (fMRI) responses elicited by emotionally evocative short videos. These stimuli were pre-annotated by human raters, who rated them according to 34 emotion categories and 14 hypothesized affective dimensions[21] (see STAR Methods for a detailed description of the emotion rating procedure; Figure 1A provides examples of emotion category ratings). Our model employed ratings of emotion categories as predictors for brain activity in a regularized linear regression, as depicted in Figures 1B and 1C. Subsequently, we utilized principal component analysis (PCA) to explore the structure and cortical representation of the underlying fundamental affective space. Our findings suggest that emotional experiences are distributed in broad, smooth gradients across large areas of the cerebral cortex. To assess whether the recovered fundamental affective space captures hypothesized affective dimensions, such as valence and arousal, we compared each of its dimensions to the 14 hypothesized affective dimensions (see STAR Methods and Figure 1D). Our results provide quantitative interpretations for the recovered dimensions and demonstrate that many hypothesized affective dimensions are indeed captured by this fundamental affective space. Our study provides an important step toward a comprehensive understanding of emotion representation in the human brain.

RESULTS

Cortical distribution and prediction performance of voxels associated with emotion

To explore the cortical distribution of voxels associated with emotion perception, we evaluated the prediction performance of the voxel-wise encoding model using the test dataset (360 samples that were not used in encoding model fitting, see STAR Methods). Prediction performance was quantified as the Pearson's correlation coefficients between the predicted and measured brain activity. To confirm that these results would not be affected by the responses evoked by visual and semantic contents, we performed banded ridge regression[22] (see STAR Methods) to explain away spurious correlations between emotion ratings and visual as well as semantic information. Figure 2A exhibits the prediction accuracy projected onto the cortical map for S03 (results for other subjects are shown in Figure S1, and we also reported the the mean absolute error (MAE) of our encoding model in Figure S2). The emotion encoding model accurately predicts brain activity in the occipitotemporal cortex, temporoparietal cortex and prefrontal cortex. On average, 21% of cortical voxels were predicted to be significant ($p < 0.01$, false discovery rate (FDR)-corrected; 26% in S01, 25% in S02, 23% in S03, 20% in S04, 12% in S05). The relatively low prediction accuracy of S05 may be due to his refusal to use a bite-bar to fix his head.[18] As shown in Figure 2B, there are many cortical regions with a higher proportion of significant predicted voxels, such as TPJ ($p < 0.01$, FDR-corrected; 62% in S01, 51% in S02, 54% in S03, 56% in S04, 43% in S05), IPL ($p < 0.01$, FDR-corrected; 50% in S01, 55% in S02, 44% in S03, 46% in S04, 17% in S05) and some higher visual areas, such as LO3 ($p < 0.01$, FDR-corrected; 51% in S01, 67% in S02, 66% in S03, 52% in S04, 47% in S05). A histogram of the prediction accuracy for all cortical voxels is shown in Figure 2C. Note that, compared with tasks such as vision,[23,24] language[25,26] and motor[27,28] associated with specific functional areas, emotion category representation is widely distributed across the entire cerebral cortex. Furthermore, we found that the predicted voxel response and measured voxel response were positively correlated even when voxel activities were averaged within each brain ROI (see Figure S3). This tendency was robustly produced in all ROIs, especially in the TPJ, PFC and LO. The above results do not support the "locationism".

Banded ridge regression allows us to investigate the brain activity encoding performance with emotional, visual and semantic features. 2D histograms displaying the comparison of model prediction accuracies using different features are shown in Figure 3A. Interestingly, the well-predicted voxels are different in the three models. To further describe these differences, we plotted two model comparisons on the same cortical map of S03 using a 2D colormap, and the results are shown in Figures 3B and 3C. On the one hand, from Figure 3B, we can find that emotion features made better encoding predictions than visual features in most brain areas beyond the primary visual cortex. On the other hand, from Figure 3C, we observe that semantic features predict better than emotion features in most higher visual cortex (e.g., V4, V7, and LO, etc), while emotion features outperformed semantic features in most temporoparietal and prefrontal cortex. These results are consistent with previous studies[18,19] using similar analysis methods.

A fundamental affective space underpinning 34 emotion categories in the brain

We employed a PCA-based analysis method to recover a fundamental affective space from the estimated weight matrix of the voxel-wise encoding model. This approach guarantees that the emotion categories represented by similar brain responses are in proximity in the estimated affective space, whereas those with significant brain response gaps are far apart.[29] To ensure the robustness of the recovered affective space, we only considered voxels that were significantly predicted ($p < 0.001$, FDR-corrected) by the encoding model. Moreover, we vertically concatenated voxel weights across five subjects to obtain group weights, which were then subjected to PCA. We have conducted the Wachter procedure[30] to the group weights and achieved quantile-quantile (Q-Q) plot of the observed singular values versus the quantiles obtained from the inverse cumulative distribution function (CDF) of the Marčenko–Pastur distribution as shown in Figure 4A. We used eventual deviations from the identity line to help finding the threshold that separates the "good" from the "unnecessary" principal components and obtained eight necessary principle components. Given the limited scope of fMRI data and finite video stimuli, we anticipated that only the first few dimensions of the recovered affective space would be meaningful and accurately reflect the true underlying affective space. Therefore, we focused our analysis on the top four PCs, whose explained variance exceeded 5% among all 34 dimensions.

The above analyses relied on the group weights obtained from all subjects. In order to examine the potential inter-subject consistency of the fundamental affective space, we computed the individual and sub-group affective spaces for each subject, where the sub-group affective space was constructed using the pooled data from the remaining four subjects. The representational similarity analysis (RSA)[31] was utilized to evaluate the cross-subject similarity between these two affective spaces. Specifically, we calculated the correlations between the

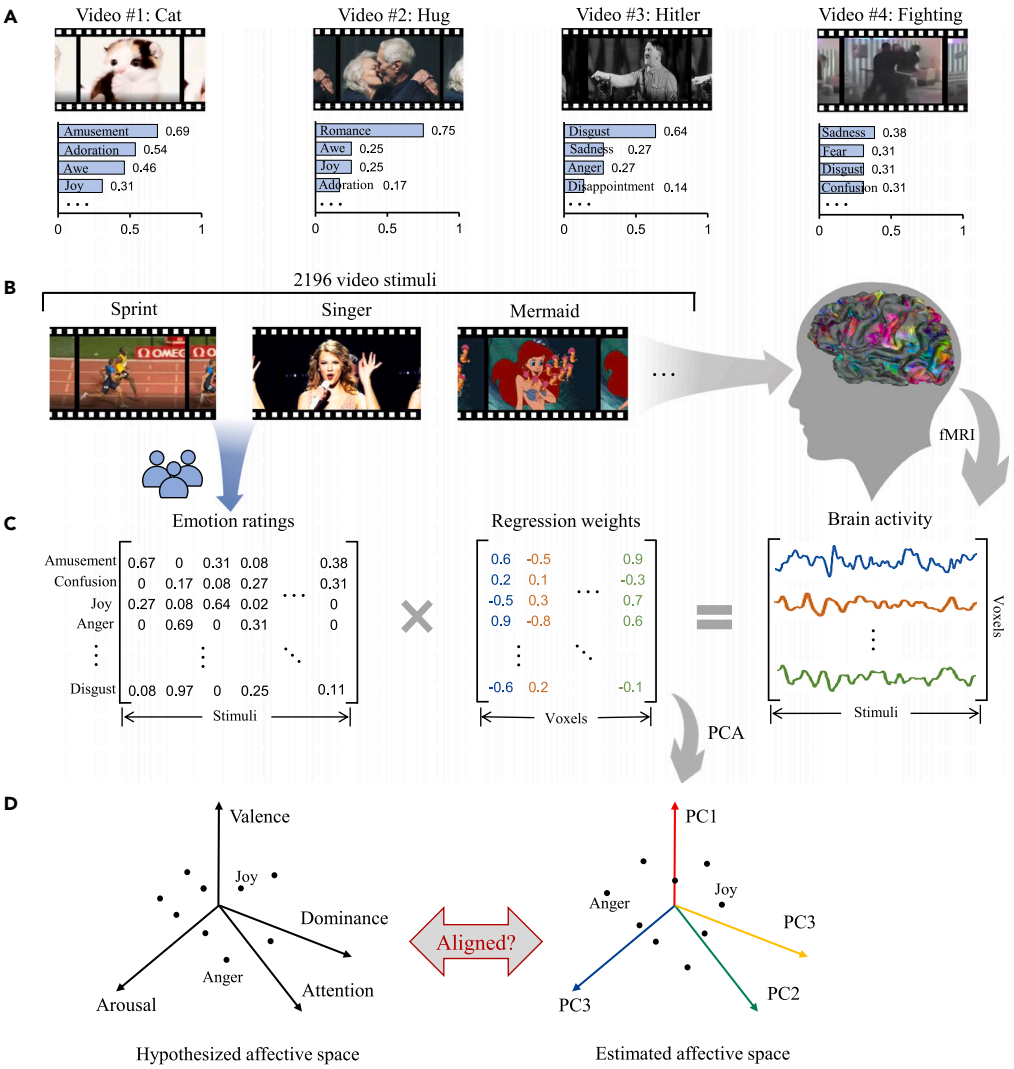


Figure 1. Schematic diagrams of the experiment and analysis methods

(A) Example screenshots of four video stimuli with corresponding emotion category ratings.

(B) A total of 2196 video stimuli were presented to five subjects while brain activity was measured using fMRI.

(C) A voxel-wise encoding model was used to predict brain activity as a linear weighted sum of the emotion category ratings, with L2-regularization.

(D) The estimated affective space was compared to hypothesized affective dimensions.

representational dissimilarity matrix (RDM) of the individual affective space and the sub-group affective space for each subject. Our results, shown in Figure 4B, indicated relatively high correlations between the individual and sub-group affective spaces for all subjects (ranging from 0.52 to 0.61, with a mean of 0.56). In comparison, we also computed the correlations between the RDM of the individual affective space and the behavioral space for each subject (ranging from 0.20 to 0.25, with a mean of 0.23). Notably, we found that the correlations between the individual and sub-group affective spaces were significantly higher than those between the individual and behavioral spaces, except for S01 (bootstrap test, $p < 0.001$). This suggests that the fundamental affective space exhibits strong inter-subject consistency, which cannot be explained by stimulus features alone.

Visualization of the fundamental affective space and its cerebral topography

To further investigate the structure of the underlying fundamental affective space, we utilized two distinct visualization techniques. Firstly, in Figure 4C, we mapped the 34 emotion categories onto the space using their PCA loadings, whereby each emotion category was assigned a unique RGB color based on the first three PCs. In this representation, emotion categories with similar representations were assigned similar colors and positioned closely within the two-dimensional space. Notably, our results diverge from a previous study[19] in which emotion categories clustered according to the polarity of emotion. We further explored this difference by constructing a behavioral

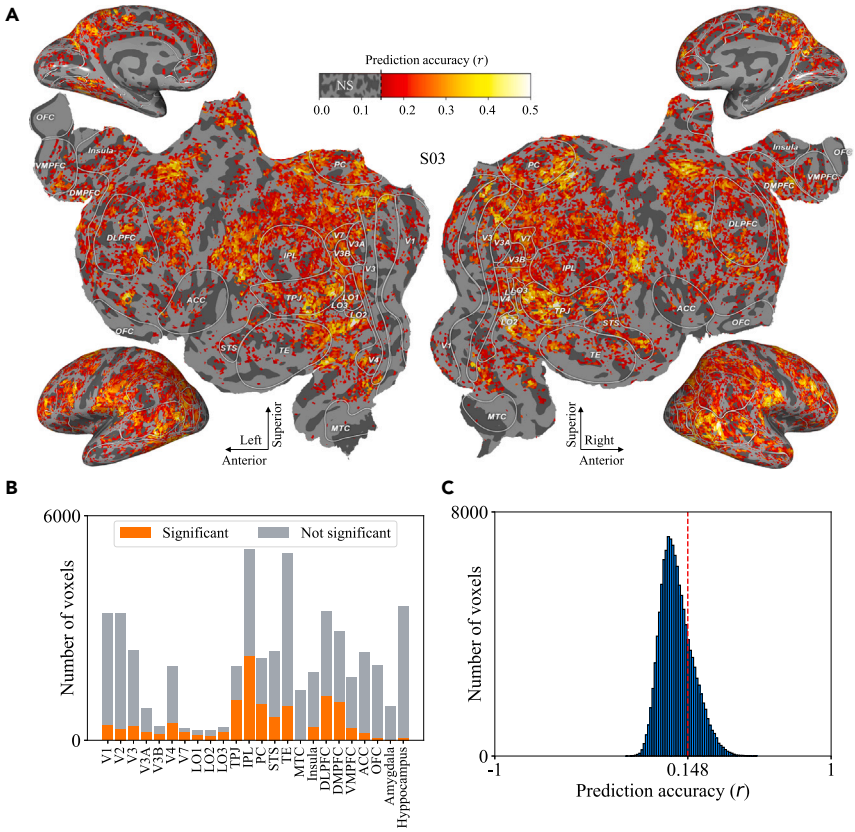


Figure 2. Model prediction performance across the cortical surface

(A) Cortical map of model prediction accuracy on both inflated and flattened cortical sheets of S03 in terms of significantly predicted voxels ($p < 0.01$, FDR-corrected), in which well-predicted voxels appear yellow.

(B) Proportion of significantly predicted voxels in representative cortical regions for S03. (abbreviations: V, visual; LO, lateral occipital; TPJ, temporo-parietal junction; IPL, inferior parietal lobule; PC, precuneus; STS, superior temporal sulcus; TE, temporal area; MTC, medial temporal cortex; DLPFC/DMPFC/VMPFC, dorsolateral/dorsomedial/ventromedial prefrontal cortex; ACC, anterior cingulate cortex; and OFC, orbitofrontal cortex).

(C) Histogram of prediction accuracy for all cortical voxels for S03. The red line indicates the threshold for significant prediction ($p < 0.01$, FDR-corrected).

data-driven affective space, based on emotion category ratings, and visualized this space in Figure 4D. Here, we observed that positive emotions were distributed in the lower left, negative emotions in the right, and ambiguous emotions in the upper left, suggesting that the emotional semantic similarity, such as polarity, can be captured by the behavioral data-driven affective space, which is consistent with a previous study.[17] These findings indicate that the neural representations of emotional experiences may not always align with the semantics of emotion categories described in natural language. Therefore, we acknowledge that the interpretation of the fundamental affective space is complex and may benefit from further exploration from new perspectives, such as linking it to the hypothesized affective dimensions (e.g., valence and arousal).

The results of the aforementioned principal component (PC) analysis indicate that the brains of different subjects represent emotion categories in a shared affective space. To investigate how this affective space is represented across the cortical surface, we assigned an RGB color to each voxel based on its projection onto the first three PCs. This allowed us to visualize how the emotion categories and PCs change over the cortical surface. In Figure 4E, we present cortical maps of the fundamental affective space on both inflated and flattened cortical sheets of S03 (see Figure S5A for other subjects). Here, RGB color is determined by voxel coefficients in the top three PCs [PC1 (red), PC2 (green), PC3 (blue)]. We can observe that the brain regions that represent emotion category in Figure 4E roughly coincide with the brain regions that have high prediction accuracy in Figure 2A, which means that the emotion category representation is consistent and stationary. Additionally, we show the projections of voxel coefficients onto PC1-PC4 for S03 in Figure 4F (and see Figure S5B for PC5-PC8), where voxels with positive projections appear red, those with negative projections appear blue.

The results suggest that emotion information is encoded in intricate patterns across almost the entire cortex, particularly in regions such as the occipitotemporal cortex, temporoparietal cortex, and prefrontal cortex, which are well-known for their involvement in emotional processing.[14,18,32] Interestingly, we observed some discrepancies between the cortical mappings in the left and right hemispheres, a phenomenon that has been reported in previous studies.[33,34] Specifically, the proportion of voxels representing emotions in the right TPJ area was higher than that in the left corresponding area, while the left TE area encoded more kinds of emotions than the right TE area. Further, the superior left

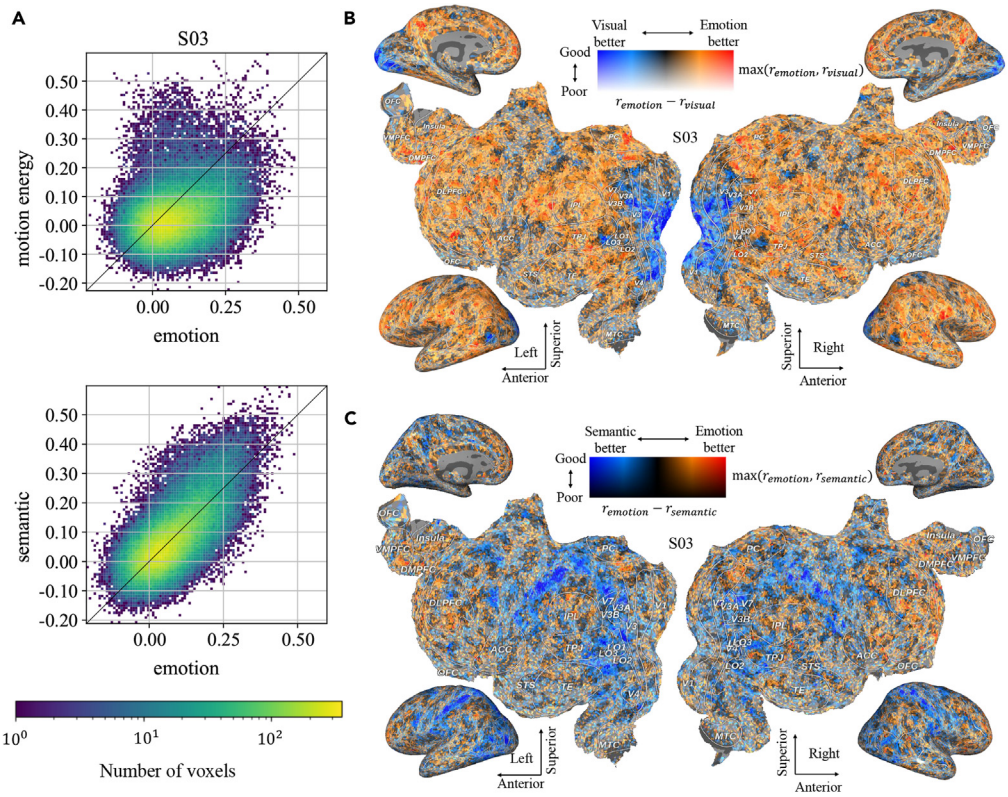


Figure 3. Disentangling the respective contributions of visual, semantic, and emotion features in voxel-wise encoding

(A) Comparison of model prediction accuracies with a 2D histogram. All voxels are represented in this histogram, where the diagonal corresponds to identical prediction accuracy for both models. A distribution deviating from the diagonal means that one model has better predictive performance than the other.

(B) Differences in prediction accuracies of emotion and visual features are projected on the cortical map, where voxels with higher prediction accuracy of visual features are shown in blue, and voxels with higher prediction accuracy of emotion features are shown in red.

(C) Emotion feature against semantic feature, where voxels with higher prediction accuracy of semantic features are shown in blue, and voxels with higher prediction accuracy of emotion features are shown in red. Results for other subjects are shown in Figure S4.

VMPFC region encodes Joy, Contempt, and Pride (Blue) while the inferior left VMPFC region is selective for Boredom, Guilt, and Envy (Purple). However, this presents the opposite representation in the right VMPFC region. Importantly, this topographical organization appears to be consistent across subjects (also see Figure S5A), suggesting that our analyses have provided a fundamental representation of the affective space in the cerebral cortex. Furthermore, the cortical maps reveal that these patterns appear to form smooth gradients across large areas of the cerebral cortex, which will be discussed further in the following section.

Interpretation of the fundamental affective space using hypothesized affective dimensions

Although previous studies have also explored the neural representation of both emotion categories and affective dimensions in the cerebral cortex,[19,35] they ignored that combining the two emotion models could achieve better interpretation of the fundamental affective semantic space. There has been work to study the relationship between emotion categories and affective dimensions from a behavioral perspective,[36] which inspired us to study the relationship between the fundamental affective space derived from the emotion encoding model and 14 hypothesized affective dimensions (e.g., valence and arousal). Our assumption is that some affective dimensions can be captured by the recovered fundamental affective space, meaning that the latter can be partially explained by the former.

To explore whether hypothesized affective dimensions can be captured by the recovered fundamental affective space, we related each of the group PCs to 14 hypothesized affective dimensions. For each affective dimension, we first assigned a 34-dimensional vector corresponding to the 34 emotion categories using L2-regularized linear regression. The sign of each element in the vector represents the affective dimension polarity of each corresponding emotion category, and the amplitude of each element denotes the affective dimension intensity, as shown in Figure S6. For example, "Sexual desire" and "Horror" are high arousal emotions that are more stimulating, while "Amusement" and "Calmness" are relatively low arousal emotions that are more subbued. We then calculated the correlation between each of the group PCs and each hypothesized affective dimension to interpret the group affective space. If a hypothesized affective dimension can effectively describe one of the group PCs, then the corresponding vector will be significantly correlated with that PC.

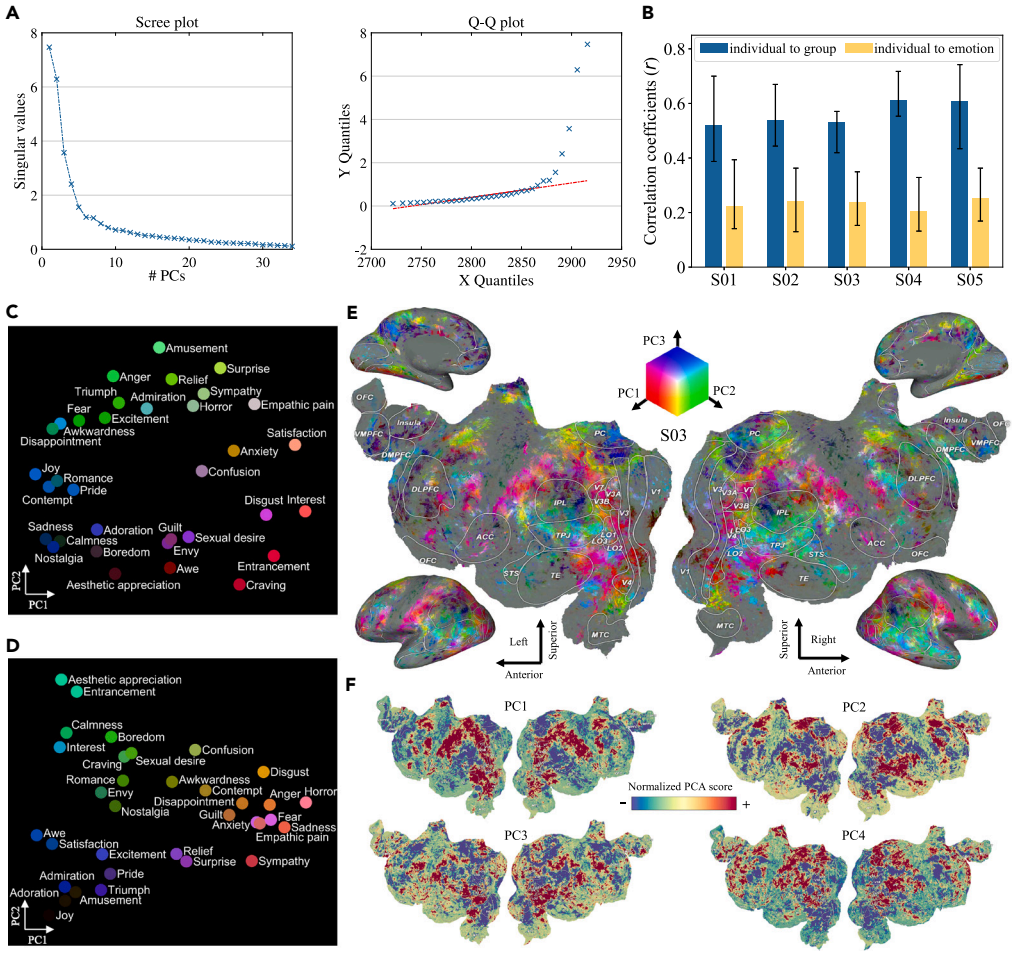


Figure 4. Fundamental affective space and cortical mapping

(A) Singluar values of covariance matrix when conducting PCA and Q-Q plot of the observed singular values versus the quantiles obtained from the inverse CDF of the Marčenko–Pastur distribution.

(B) Cross subject consistency of the fundamental affective space. The blue bar indicates individual to group correlations which was calculated with individual affective space and sub-group affective space for each subject. The yellow bar indicates individual to emotion correlations which was calculated with individual affective space for each subject and behavioral semantic space (see STAR Methods). Error bars show the minimum and maximum correlation coefficients over the 1000 bootstrap samples ($p < 0.001$).

(C) Affective space constructed from brain activity. The 34 emotion categories were organized based on their coefficients on the first and second PCs. The color of each marker is determined by an RGB colormap based on the category coefficients in the top three PCs. The position of each marker is determined by the coefficients of PC1 and PC2. This ensures that categories that are represented similarly in the brain appear near each other.

(D) Affective space constructed from behavioral data.

(E) Cortical maps of the fundamental affective space on both inflated and flattened cortical sheets of S03. RGB Color is determined by the voxel coefficients in the top three PCs.

(F) Projection of voxel coefficients onto the individual PC for S03.

The correlations between the top four group PCs and 14 affective dimensions are shown in Figure 5A (panoramic view of the correlations between all 34 PCs and the 14 hypothesized affective dimensions is presented in Figure S7). We can observe that affective dimensions are not exactly aligned with group PCs, and each of the group PCs captures more than one affective dimension. The first PC is best explained by arousal ($r = 0.30, p < 0.05$; ranging from more subdued to more stimulated) and is also well explained by identity ($r = -0.29, p < 0.05$; ranging from strong of group identity to lack group identity). The second PC is best explained by approach ($r = -0.56, p < 0.05$; ranging from desire to approach to desire to avoid) and is also well explained by control ($r = -0.50, p < 0.05$; ranging from under control to out of control). The third PC is best explained by attention ($r = -0.62, p < 0.05$; ranging from more focused to unfocused) and is also well explained by safety ($r = 0.55, p < 0.05$; ranging from unsafe to safe). The fourth PC is best explained by commitment ($r = 0.34, p < 0.05$; ranging from lack of commitment to strong commitment to an individual or creature) and is also well explained by safety ($r = -0.32, p < 0.05$; ranging from safe to unsafe). It should be noted that a negative correlation between a group PC and a hypothesized affective dimension indicates that the

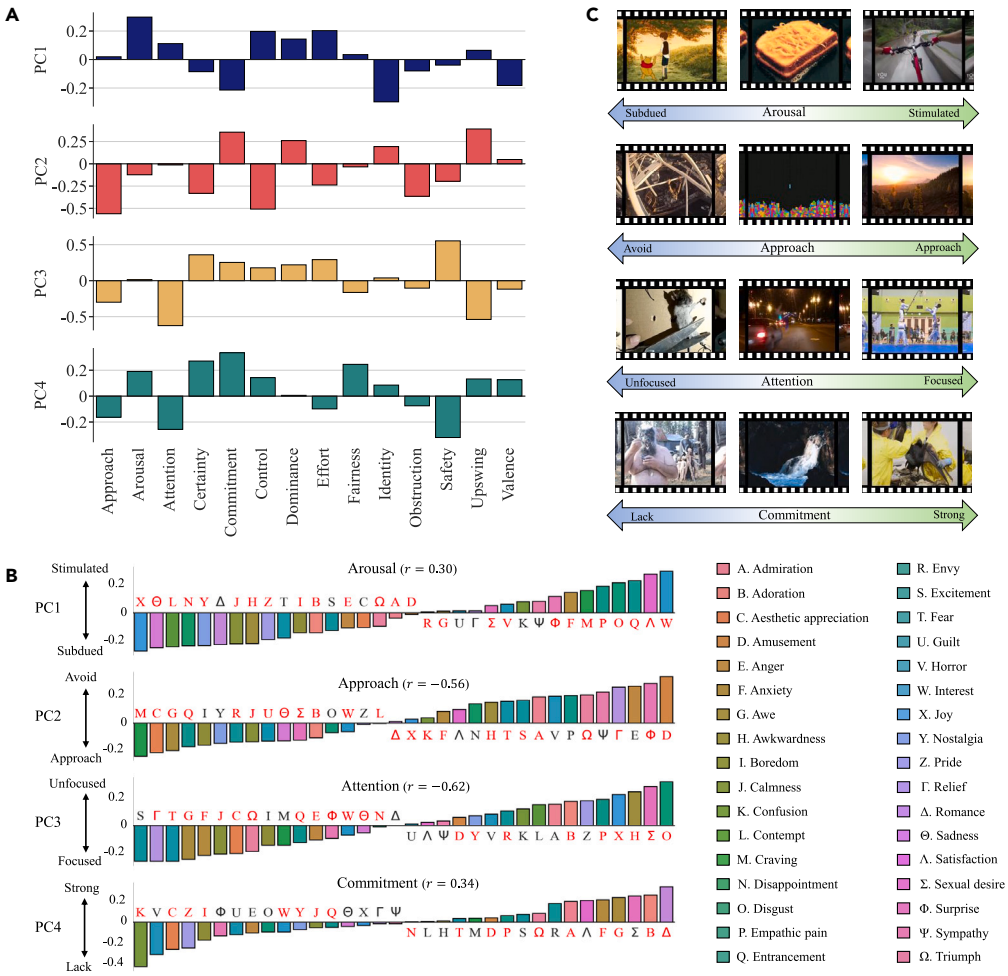


Figure 5. Interpretation of the recovered fundamental affective space

(A) Pearson's correlation coefficients between the top four PCs and 14 hypothesized affective dimensions are compared.

(B) The emotion category projections corresponding to the top four PCs. For each PC, the most related affective dimension is shown at the top: arousal for PC1 ($r = 0.30$), approach for PC2 (negative correlation, $r = -0.56$), attention for PC3 (negative correlation, $r = -0.62$), and commitment for PC4 ($r = 0.34$). Emotion categories that match with the PC's best explained affective dimension in terms of polarity are marked in red.

(C) Examples of video screenshots for arousal, approach, attention, and commitment according to the polarity of them measured by a 9-point Likert scale.[21]

direction of the PC is opposite to the polarity of the dimension. These findings offer quantitative interpretations of the top four group PCs in relation to the hypothesized affective dimensions. However, it is important to note that many affective dimensions are not fully captured by the top four group PCs, and it is possible that they may be better represented by lower-variance group PCs that were not significantly discernible in our analysis. Surprisingly, valence, which represents the polarity of emotion, as a common affective dimension has not been encoded by these higher-variance group PCs. In order to explore this phenomenon, we conducted variance inflation factor (VIF) test for the multi-collinearity of these affective dimensions. The results shown in Table S1 indicate that valence has the largest VIF value (9.549), which means that there exists a significant collinearity between valence and other affective dimensions. Furthermore, we have exhibited the correlation map between these affective dimensions in Figure S8A, which illustrates that valence has strong linear correlation with other dimensions, especially with approach ($r = 0.9$). This also provides a possible explanation for valence not being captured. Furthermore, we conducted an RSA on two types of affective spaces: one constructed from behavioral data (i.e., 14 affective dimensions), and the other derived from the estimated PCA space (a total of 34 dimensions). The results, as shown in Figure S8B, indicate a strong correlation between the two spaces (average correlation $\rho = 0.29$, $p < 0.001$).

To further explore how well the group PCs can be explained by 14 hypothesized affective dimensions, we visualized the PCA loadings of the top four group PCs and attempted to explain each of them in terms of the most related affective dimension, as shown in Figure 5B. We sorted the emotion category projections of each PC and marked the emotion categories that matched with the PC's best explained affective dimension in terms of polarity. Our results demonstrated that PC1 was associated with arousal and matched 26 emotion categories, PC2 was associated with approach and matched 24 emotion categories (with a negative correlation), PC3 was associated with attention and matched

22 emotion categories (with a negative correlation), and PC4 was associated with commitment and matched 18 emotion categories. All four group PCs matched more than half of the emotion categories with their best explained affective dimensions, indicating that the hypothesized affective dimensions could reasonably explain the top four PCs.

Furthermore, we provided video screenshots that represented different levels of arousal, approach, attention, and commitment in Figure 5C according to their polarity. For instance, videos with high arousal were associated with heterosexual behavior, while videos with low arousal were associated with same-sex sexual behavior. As a result, even in the affective dimension vector corresponding to arousal, the value of Sexual desire was the largest, and Sexual desire (Σ) was closer to 0 in the PCA loadings of PC1. In addition, some cartoons such as "*Winnie the Pooh*" have relatively low arousal, and emotions such as Joy (X) and Calmness (J) they convey also have relatively low PCA loadings in PC1. Videos such as bike races have relatively high arousal, and the Interest (W) and Entrancement (Q) in which they evoke also have relatively high PCA loadings in PC1. For approach, videos such as extreme sports have a relatively low approach. Even if most people desire to avoid participating in this kind of sport, we still feel Amusement (D) and Surprise (Φ) as viewers. These emotions have relatively high PCA loadings in PC2 (negative correlation). Videos such as landscape have a relatively high approach, and they evoke Craving, Asethetic appreciation and Awe, which have relatively low PCA loadings in PC2 (negative correlation). For attention, videos associated with sex have relatively low attention, which means they are more unfocused. Disgust (O), Sexual desire (Σ) and Awkwardness (H) are the emotions that these videos mainly convey and captured in high PCA loadings in PC3 (negative correlation). On the other hand, videos such as Taekwondo make viewers more focused and therefore have relatively high attention. Fear (T), Awe (G) and Anxiety (F) are the emotions that these videos mainly convey and captured in low PCA loadings in PC3 (negative correlation). For commitment, which is an affective dimension related to individuals or creatures, the scene including a man wearing a mask made of animal hair apparently has relatively low commitment, while the scene of rescuing animals has relatively high commitment. Confusion (K) and Boredom (I) conveyed by the former have low PCA loadings in PC4, while Romance (Δ), Adoration B. and Awe (G) conveyed by the latter have high PCA loadings in PC4. These results illustrate that interpretation of group PCs using hypothesized affective dimensions is reasonable and reliable to a certain extent.

Smoothness and gradients of cortical maps

The preceding analysis has revealed that the estimated affective space used to support emotion categories exhibits significant inter-subject consistency. Additionally, the affective maps displayed in Figure 4E exhibit smoothness across extensive regions of the cerebral cortex. However, these results alone are insufficient to demonstrate that the apparent smoothness of the cortical map is a property that is unique to the estimated affective space. Therefore, we must assess the possibility that any 4-D projection of the voxel weights onto the cortex will produce a smooth mapping.

To address this issue, we first projected the encoding model weights into the estimated affective space and randomly constructed space, respectively. The qualitative results are depicted in Figure 6A, and it is evident that the affective map in the random space does not possess a smooth spatial distribution across the cortex, as opposed to the map displayed in Figure 4E. To explore this smoothness qualitatively, we calculated the correlation between the projections for each pair of voxels using for constructing affective space ($p < 0.001$, FDR-corrected). Then, we collected and averaged these pairwise correlations on account of the distance between each pair of voxels. For estimating the null distribution of smoothness values and establish statistical significance, we repeated this procedure using 100 random semantic spaces (see STAR Methods for details). Figure 6B exhibits the mean correlation between voxel projections into the estimated affective space and randomly constructed space as a function of the distance between voxels. Results show that the mean correlations of the estimated affective space projections are significantly greater than chance ($p < 0.01$) in all subjects but S05 (perhaps attribute to the poor data quality for his refusal to use a bite-bar to fix his head) for adjacent voxels (distance 1) and voxels separated by one intermediated voxel (distance 2). Taken together, these findings imply that the smoothness of the cortical map is a unique feature of the estimated affective space, lending further support to the significance of the estimated affective space.

Cortical maps of the estimated affective space appear to form gradient-like topographic organization in both hemispheres, and these gradients appear to be distributed across large areas of the cerebral cortex. We schematically portrayed four gradients corresponding to the estimated affective space (indicated by white arrows and numbers) for S03 in Figure 7. The results demonstrated that gradient 1 starts in the posterior inferior parietal lobule (IPL) and shifts toward the anterior cingulate cortex (ACC), gradient 2 starts in the posterior temporo-parietal junction (TPJ) and ends in the inferior temporal area (TE). These two gradients observed from PC1 are anchored on one end by unimodal areas and on the other end by transmodal areas (IPL and TPJ), which are related to the arousal dimension demonstrated above corresponding to the transition from subdued to stimulated emotions. Gradient 3 starts in the parietooccipital sulcus then crosses the precuneus (PC) and ends in the posterior paracentral lobule. This gradient is depicted in PC2 and related to approach dimension accounting for the transition from avoidance to approach. Gradient 4 starts in the posterior dorsolateral prefrontal cortex (DLPFC) and shifts toward dorsomedial prefrontal cortex (DMPFC). This gradient is located in prefrontal cortex and can be captured by PC3 (from unfocused to focused) mostly and PC4 (from lack to strong commitment) marginally. We notice that most of these gradients are distributed over the default mode network (TPJ, IPL, PC and DMPFC) and exhibit smooth transitions in terms of colors. Furthermore, comparing Figures 7 with S5A, we found that the portrayed affective gradients are consistent across subjects.

DISCUSSION

We employed a voxel-wise encoding model[20] to investigate the representation of diverse emotion categories in the human cerebral cortex using fMRI data obtained from a range of short videos. Our results revealed that the representation of emotion categories is not localized in

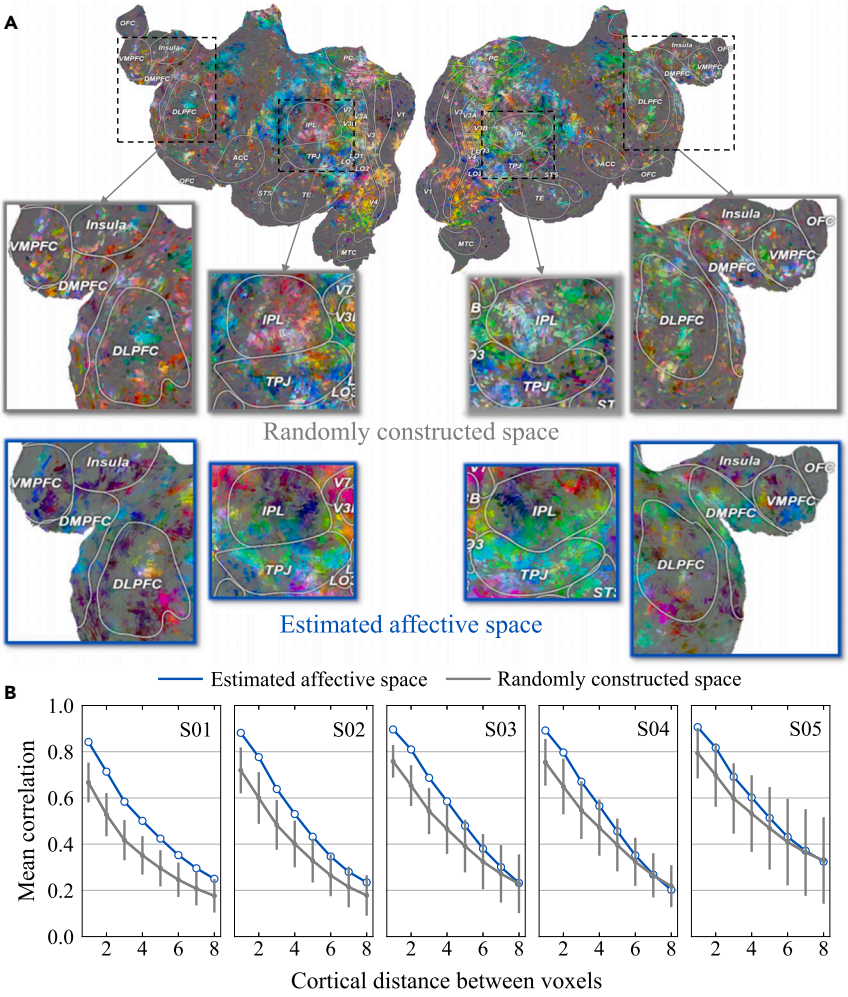


Figure 6. Smoothness of the cortical maps

(A) Qualitative comparison of randomly constructed space (gray) and the estimated affective space (blue) for typical subject (S03).

(B) Quantitative comparison of smoothness. Gray error bars show 95% confidence intervals for the random space results. For adjacent voxels (distance 1) and voxels separated by one intermediated voxel (distance 2), smoothness of estimated affective space projections are significantly greater than chance ($p < 0.01$) in all subjects but S05.

discrete cortical areas, but rather distributed and overlapping patterns in transmodal and prefrontal brain regions. Additionally, the cortical maps derived from our affective space estimation were significantly smoother than expected by chance, supporting the notion that emotional experiences are organized along smooth gradients that are distributed topographically across the cortex. These findings provide compelling evidence that challenges the notion of discrete and localized cortical representations of emotions.

The principle of gradient organization has been demonstrated to be a fundamental organizing principle for the brain to effectively represent and integrate information from external stimuli in numerous fMRI studies. For example, in the primary visual cortex, two orthogonal and spatially overlapping gradients, visual eccentricity and angle selectivity, form retinotopy.[37,38] Recently, a study on the neural representation of emotions revealed that three orthogonal and spatially overlapping gradients encode the polarity, complexity, and intensity of emotional experiences in the right TPJ, which is referred to as "emotionotopy".[17] Unlike retinotopy, emotionotopy is not predetermined but discovered through data-driven approaches. Moreover, several previous studies have demonstrated that the principle of gradient organization applies to the representation of other higher-order information.[29,39–44] Our findings are consistent with these studies and further support the idea that emotional states are gradient-like encoded in the human brain.

Previous studies have demonstrated that emotionally charged videos can elicit significant brain activity in various brain regions and have made notable contributions to the understanding of emotional functions.[16,18,45] Nevertheless, these studies did not aim to systematically map the representation of emotions or to discover the underlying affective space. Our findings shed light on why affective videos can elicit consistent neural activity across different subjects: emotion categories are represented based on a common affective space that is consistently mapped to cortical anatomy.

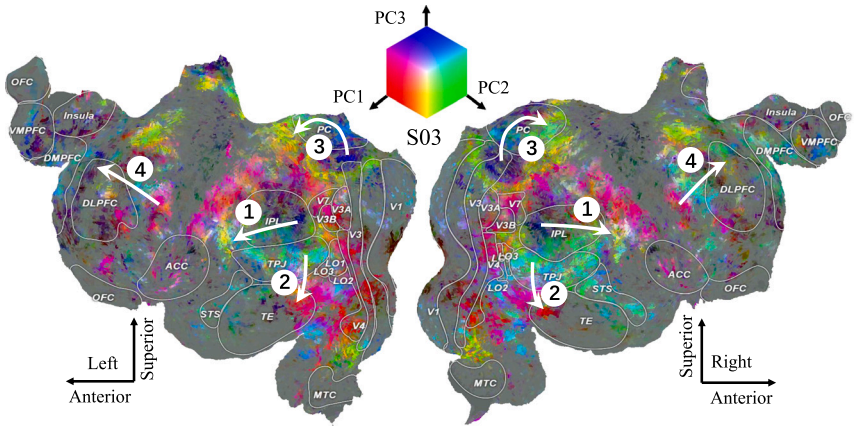


Figure 7. Affective gradients on the cortical maps
Gradients corresponding to PC1-PC4 were indicated by white arrows and numbers.

There has also been a persistent and controversial debate in affective sciences which focused on whether emotions are better conceptualized as emotion categories or affective dimensions. A recent study[18] investigated this issue with mapping the brain activity to emotion category or affective dimension ratings directly and arrived at the conclusion that the cortical representation of emotion is categorical. However, it is well-known that affective dimension ratings are noisier than categorical reports and a direct comparison between the two models could favor the latter simply because of this SNR difference. In our opinion, combining the two models has more potential to settle the debate. There existed a study which reconciled the issue of emotion category and affective dimension from the perspective of time dynamics.[3] They came to the conclusion that emotion category prevails in perceptual and early frontotemporal cortex and that affective dimension impinges predominantly on a later limbic–temporal network. However, they only focused on five basic emotion category and used synthetic voice data as stimuli. Compared with this study, we utilized video stimuli of real scenes and considered more fine-grained emotion categories, which supports us to unify the two theory from the point of semantic space. This is a more promising way to study the emotion representation in human brain, especially as the number of emotion categories is growing with the development of psychology.

A recent study suggests that human visual cortex encodes emotion-related information and can support decoding multiple categories of emotional experiences.[35] In this study, however, we found that emotion-related representations were not predominantly distributed in the visual cortex but in several transmodal and frontal regions (IPL, TPJ, TE, DLPFC, DMPFC, and VMPFC) traditionally known to be associated with emotional processing.[14,18,32] We identified many transmodal, prefrontal, and lateral occipital regions (IPL, TPJ, TE, DLPFC, DMPFC, VMPFC, LO1, LO2, and LO3) as responsive to affective videos, while the same voxel-wise encoding analysis failed to reveal the subcortical region amygdala traditionally associated with emotional processing. This result aligns with a previous study[19] using affective videos and voxel-wise encoding analysis. Additional analyses of the subcortical regions (amygdala, brainstem, caudate, cerebellum, pallidum, putamen, hippocampus, hypothalamus, thalamus, and nucleus accumbens) showed that affective videos increased activation intensity in many of these regions, however, most voxels in these regions did not survive statistical thresholds ($p < 0.01$, FDR-corrected) in a whole-brain analysis. Therefore, in this work, we focused on the cortex. Furthermore, we found that semantic features appeared to have stronger predictive power for neural activity in the lateral occipital cortex (LOC) than emotion features, but this does not mean that LO regions are unrelated to emotion processing. Notably, LOC has also been consistently identified in neuroimaging studies using affective visual stimuli[45,46](but not in other modalities[2,47,48]).

Our findings are in line with previous studies[17,19] and support the notion of a lower-dimensional, biologically plausible affective space for representing emotions in the cerebral cortex. The gradient-like organization of emotional states in this affective space is consistent with the "constructionism" perspective of emotion coding, suggesting that the spatial arrangement of distinct gradients allows the brain to efficiently represent a wide range of emotional states within a limited number of brain regions. Collectively, our study represents a significant advancement in our understanding of how emotions are represented in the human brain.

Limitations of the study

The limitations of the present study are as follows. First, the present study used affective video clips to mimic real-life emotional experiences from natural events or social interactions. Nevertheless, the emotional experiences evoked by these stimulus video clips may differ from real-life natural events or social interactions in their magnitude, occurrence rate and co-occurrence. Second, the emotion ratings used to construct the encoding model may have biased the recovered affective space. Specifically, the emotion ratings used in this study were annotated by third-party participants who were independent of the fMRI subjects. Due to individual differences in emotional experiences (e.g., for the same video, some people may be surprised while others are not), this emotion rating way may introduce some bias slightly, as third-party participants may experience emotions differently than fMRI subjects. Third, despite controlling for visual and semantic features in the voxel-wise

encoding analysis (by banded ridge regression;[22] see STAR Methods), it is possible that such features still confound the results. Future studies need to design better paradigms to rule out the potential contamination of visual and semantic features on emotion representation.

Surprisingly, we have observed that the prediction performance of the encoding model is relatively low in vmPFC, which has been proved to be associated with human affective experience.[49] Coincidentally, previous study[49] has shown that vmPFC responses are particularly variable across individuals leading to generally lower decoding accuracy using pattern classification as well. We also found such individual differences, for instance, the prediction accuracy of S01 was relatively higher than other subjects. Existing study[49] has pointed out that aligning vmPFC responses across individual requires developing a new approach that detects changes in latent states. Limited by the length of the video stimuli (minimum 0.15 s), we did not consider temporal information in our study, but averaged the fMRI over time for each stimulus. In the future, using longer video stimuli to construct a dynamic affective space may have more potential to explain the behavior of vmPFC.

In addition, the number of videos used for regression was unbalanced across emotion categories, which may impact the encoding model fitting. In estimating the relationship between emotion categories and affective dimensions, some categories with a large number of samples were associated with relatively high weights on every affective dimension which may be due to the large number of samples in those categories and may not reflect the actual relationship between emotion categories and affective dimensions. Fortunately, most categories that appeared neither too frequent nor too rare in these stimuli are largely immune to this bias. Therefore, we do not believe that these biases have a significant impact on the results of this study.

The fMRI dataset utilized in our study comprised only five subjects. Due to the complexity and dynamic nature of the brain, a small sample size may limit the efficiency of the dataset, and the obtained voxel-wise encoding model may lack robustness when applied to real-world scenarios. Additionally, the number of video stimuli used in this study was not extensive enough. As the accuracy of the encoding model fitting relies on the availability of a sufficient number of paired 'stimuli-response' data, insufficient paired data may result in a deviation of the fitted model from reality. Therefore, to improve the generalizability of the encoding model, future studies should consider recruiting a larger sample size and building a more comprehensive 'stimuli-response' dataset.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Subjects
  - Video stimuli
  - Emotion ratings of stimuli
  - fMRI experimental paradigm
  - fMRI data acquisition
  - fMRI data preprocessing
  - Flatmap construction
  - Eliminating spurious correlations between emotion ratings and other related information
  - Voxel-wise encoding model fitting and testing
  - Principal component analysis
  - Cross-subject consistency of the fundamental affective space
  - Comparison between recovered and hypothesized affective dimensions
  - Smoothness of cortical maps under estimated affective space
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107571.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China 2022ZD0116500; in part by the National Natural Science Foundation of China under Grant 62206284 and Grant 61976209; in part by the Beijing Advanced Discipline Fund; and in part by the CAAI-Huawei MindSpore Open Fund. We would like to thank Alan S. Cowen for sharing the video stimuli and emotion rating data. We also thank Tomoyasu Horikawa and Yukiyasu Kamitani for sharing the evoked fMRI data. We also thank Luca Cecchetti (a non-anonymous reviewer) for his helpful comments that have improved this paper.

AUTHOR CONTRIBUTIONS

C.D., K.F., and H.H. designed the research; K.F. conducted the experiments; C.D., K.F., and B.W. analyzed the results; C.D. and K.F. wrote the paper. All the authors proof and approve the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 10, 2023
Revised: July 3, 2023
Accepted: August 7, 2023
Published: August 12, 2023

REFERENCES

1. Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends Cogn. Sci. *16*, 458–466.
2. Chikazoe, J., Lee, D.H., Kriegeskorte, N., and Anderson, A.K. (2014). Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. *17*, 1114–1122.
3. Giordano, B.L., Whiting, C., Kriegeskorte, N., Kotz, S.A., Gross, J., and Belin, P. (2021). The representational dynamics of perceived voice emotions evolve from categories to dimensions. Nat. Hum. Behav. *5*, 1203–1213.
4. Vytal, K., and Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J. Cogn. Neurosci. *22*, 2864–2885.
5. Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature *372*, 669–672.
6. Broks, P., Young, A.W., Maratos, E.J., Coffey, P.J., Calder, A.J., Isaac, C.L., Mayes, A.R., Hodges, J.R., Montaldi, D., Cezayirli, E., et al. (1998). Face processing impairments after encephalitis: amygdala damage and recognition of fear. Neuropsychologia *36*, 59–70.
7. Sprengelmeyer, R., Young, A.W., Schroeder, U., Grossenbacher, P.G., Federlein, J., Büttner, T., and Przuntek, H. (1999). Knowing no fear. Proc. Biol. Sci. *266*, 2451–2456.
8. Lettieri, G., Handjaras, G., Setti, F., Cappello, E.M., Bruno, V., Diano, M., Leo, A., Ricciardi, E., Pietrini, P., and Cecchetti, L. (2022). Default and control network connectivity dynamics track the stream of affect at multiple timescales. Soc. Cogn. Affect. Neurosci. *17*, 461–469.
9. Sachs, M.E., Habibi, A., Damasio, A., and Kaplan, J.T. (2020). Dynamic intersubject neural synchronization reflects affective responses to sad music. Neuroimage *218*, 116512.
10. Raz, G., Touroutoglou, A., Wilson-Mendenhall, C., Gilam, G., Lin, T., Gonen, T., Jacob, Y., Atzil, S., Admon, R., Bleich-Cohen, M., et al. (2016). Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences. Cogn. Affect. Behav. Neurosci. *16*, 709–723.
11. Russell, J.A. (2003). Core affect and the psychological construction of emotion. Psychol. Rev. *110*, 145–172.
12. Wager, T.D., Kang, J., Johnson, T.D., Nichols, T.E., Satpute, A.B., and Barrett, L.F. (2015). A Bayesian model of category-specific emotional brain responses. PLoS Comput. Biol. *11*, e1004066.
13. Barrett, L.F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci. *12*, 1–23.
14. Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K., and Wager, T.D. (2008). Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage *42*, 998–1031.
15. Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., and Barrett, L.F. (2012). The brain basis of emotion: a meta-analytic review. Behav. Brain Sci. *35*, 121–143.
16. Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I.P., Lampinen, J., Vuilleumier, P., Hari, R., Sams, M., and Nummenmaa, L. (2016). Discrete neural signatures of basic emotions. Cereb. Cortex *26*, 2563–2573.
17. Lettieri, G., Handjaras, G., Ricciardi, E., Leo, A., Papale, P., Betta, M., Pietrini, P., and Cecchetti, L. (2019). Emotionotopy in the human right temporo-parietal cortex. Nat. Commun. *10*, 5568.
18. Horikawa, T., Cowen, A.S., Keltner, D., and Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. iScience *23*, 101060.
19. Koide-Majima, N., Nakai, T., and Nishimoto, S. (2020). Distinct dimensions of emotion in the human brain and their representation on the cortical surface. Neuroimage *222*, 117258.
20. Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400–410.
21. Cowen, A.S., and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. USA *114*, E7900–E7909.
22. Nunez-Elizalde, A.O., Huth, A.G., and Gallant, J.L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage *197*, 482–492.
23. Ungerleider, L.G., and Haxby, J.V. (1994). 'what' and 'where'in the human brain. Curr. Opin. Neurobiol. *4*, 157–165.
24. Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. *8*, 679–685.
25. Binder, J.R., Frost, J.A., Hammeke, T.A., Cox, R.W., Rao, S.M., and Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. J. Neurosci. *17*, 353–362.
26. Catani, M., Allin, M.P.G., Husain, M., Pugliese, L., Mesulam, M.M., Murray, R.M., and Jones, D.K. (2007). Symmetries in human brain language pathways correlate with verbal recall. Proc. Natl. Acad. Sci. USA *104*, 17163–17168.
27. Fink, G.R., Frackowiak, R.S., Pietrzyk, U., and Passingham, R.E. (1997). Multiple nonprimary motor areas in the human cortex. J. Neurophysiol. *77*, 2164–2174.
28. Amiez, C., and Petrides, M. (2014). Neuroimaging evidence of the anatomo-functional organization of the human cingulate motor areas. Cereb. Cortex *24*, 563–578.
29. Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.
30. Wachter, K.W. (1976). Probability plotting of multiple discriminant ratios. Proc. Soc. Stat. Sect. Am. Stat. Assoc. Part II, 830–833.
31. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4–28.
32. Skerry, A.E., and Saxe, R. (2014). A common neural code for perceived and inferred emotion. J. Neurosci. *34*, 15997–16008.
33. Dimond, S.J., Farrington, L., and Johnson, P. (1976). Differing emotional response from right and left hemispheres. Nature *261*, 690–692.
34. Davidson, R.J. (1992). Anterior cerebral asymmetry and the nature of emotion. Brain Cogn. *20*, 125–151.
35. Kragel, P.A., Reddan, M.C., LaBar, K.S., and Wager, T.D. (2019). Emotion schemas are embedded in the human visual system. Sci. Adv. *5*, eaaw4358.
36. Russell, J.A., and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. J. Res. Pers. *11*, 273–294.
37. Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science *268*, 889–893.
38. Engel, S.A., Glover, G.H., and Wandell, B.A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex *7*, 181–192.
39. Hansen, K.A., Kay, K.N., and Gallant, J.L. (2007). Topographic organization in and near human visual area v4. J. Neurosci. *27*, 11896–11911.

40. Harvey, B.M., Klein, B.P., Petridou, N., and Dumoulin, S.O. (2013). Topographic representation of numerosity in the human parietal cortex. Science *341*, 1123–1126.
41. Sha, L., Haxby, J.V., Abdi, H., Guntupalli, J.S., Oosterhof, N.N., Halchenko, Y.O., and Connolly, A.C. (2015). The animacy continuum in the human ventral vision pathway. J. Cogn. Neurosci. *27*, 665–678.
42. Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., and Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458.
43. Margulies, D.S., Ghosh, S.S., Goulas, A., Falkiewicz, M., Huntenburg, J.M., Langs, G., Bezgin, G., Eickhoff, S.B., Castellanos, F.X., Petrides, M., et al. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. Proc. Natl. Acad. Sci. USA *113*, 12574–12579.
44. Huntenburg, J.M., Bazin, P.-L., and Margulies, D.S. (2018). Large-scale gradients in human cortical organization. Trends Cogn. Sci. *22*, 21–31.
45. Chan, H.-Y., Smidts, A., Schoots, V.C., Sanfey, A.G., and Boksem, M.A.S. (2020). Decoding dynamic affective responses to naturalistic videos with shared neural patterns. Neuroimage *216*, 116618.
46. Nielen, M.M.A., Heslenfeld, D.J., Heinen, K., Van Strien, J.W., Witter, M.P., Jonker, C., and Veltman, D.J. (2009). Distinct brain systems underlie the processing of valence and arousal of affective pictures. Brain Cogn. *71*, 387–396.
47. Kim, H.-C., Bandettini, P.A., and Lee, J.-H. (2019). Deep neural network predicts emotional responses of the human brain from functional magnetic resonance imaging. Neuroimage *186*, 607–627.
48. Putkinen, V., Nazari-Farsani, S., Seppälä, K., Karjalainen, T., Sun, L., Karlsson, H.K., Hudson, M., Heikkilä, T.T., Hirvonen, J., and Nummenmaa, L. (2021). Decoding music-evoked emotions in the auditory and motor cortex. Cereb. Cortex *31*, 2549–2560.
49. Chang, L.J., Jolly, E., Cheong, J.H., Rapuano, K.M., Greenstein, N., Chen, P.-H.A., and Manning, J.R. (2021). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. Sci. Adv. *7*, eabf7129.
50. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111–116.
51. Fischl, B. (2012). Freesurfer. Neuroimage *62*, 774–781.
52. Gao, J.S., Huth, A.G., Lescroart, M.D., and Gallant, J.L. (2015). Pycortex: an interactive surface visualizer for fMRI. Front. Neuroinform. *9*, 23.
53. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. *21*, 1641–1646.
54. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Video stimuli	Cowenlab	https://goo.gl/forms/XErJw9sBeyuOyp5Q2
Motion energy features	Open Science Framework	https://osf.io/9uyn2/
Raw fMRI data	OpenNeuro	https://openneuro.org/datasets/ds002425
Emotional ratings and preprocessed fMRI data	Figshare	https://doi.org/10.6084/m9.figshare.11988351.v1
Software and algorithms		
Custom code	Open Science Framework	https://osf.io/9uyn2/
Matlab	Mathworks	https://www.mathworks.com/
Python	Python Software Foundation	https://www.python.org/
Pycortex	Gallantlab	https://github.com/gallantlab/pycortex
Himalaya toolbox	Gallantlab	https://github.com/gallantlab/himalaya
Pymoten	Gallantlab	https://github.com/gallantlab/pymoten

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Prof. Huiguang He (huiguang.he@ia.ac.cn).

Materials availability

This study did not generated new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The video stimuli are available at https://goo.gl/forms/XErJw9sBeyuOyp5Q2. The motion energy features of the video stimuli are available at https://osf.io/9uyn2/. Raw fMRI are available at https://openneuro.org/datasets/ds002425. The emotional ratings and semantic features of the video stimuli as well as the preprocessed fMRI data are available at https://doi.org/10.6084/m9.figshare.11988351.v1.
- All original code has been deposited at https://osf.io/9uyn2/ and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This work has not involved the use of human subjects or samples, nor has it use experimental models that require reporting of experimental model and subject details.

METHOD DETAILS

We reanalyzed two existing datasets that were published previously and publicly available.[18,21] Below, we will give a brief overview of the data we used.

Subjects

Five healthy subjects (1 female; M_{age} = 25.8; range 23-34 years; referred to as S01-S05) with normal or corrected-to-normal vision participated in the visually evoked emotional experiment. All subjects signed written informed consent, and the study protocol was approved by the Ethics Committee of ATR.[18]

Video stimuli

The stimuli contained a total of 2196 affective videos collected in an earlier study.[21] The videos ranged in length from ~0.15 s to ~ 90 s. Each of the videos was resized so that both the width and height of videos were limited to 12 degrees and was presented at the center of screen on a gray background (the sound of the videos was removed, leaving only the visual stimuli).

Emotion ratings of stimuli

A total of 853 English-speaking US annotators who did not experience the fMRI experiments were recruited.[21] They were presented with 2196 emotionally evocative videos on Amazon Mechanical Turk to obtain emotion judgments regarding emotion categories and affective dimensions. For the emotion categories, annotators rated each video on 34 emotion categories (100-point scale) based on their subjective feelings. Then the ratings were binarized by converting ratings greater than 0 to 1 leading to a dichotomous yes/no response for an emotion category of a video from an individual rater. For the affective dimensions, annotators rated each video on 14 affective dimensions. The ratings were obtained on a 9-point Likert scale. Repeated emotion judgments were obtained from multiple raters (9-17), and the ratings were averaged to obtain an average score for each video on both emotion judgments.

fMRI experimental paradigm

Video stimuli were projected onto the screen of an MRI scanner, and subjects were asked to hold their heads still. To allow subjects to freely focus on the details of the stimuli, subjects were allowed to watch video stimuli without fixation. The experiment contained 61 separate runs.[18] Each run consisted of 36 stimulus blocks whose durations varied according to the durations of videos presented in each stimulus block. For stimulus blocks with video shorter than 8 s, researchers presented the same video stimulus repeatedly until the total presentation duration exceeded 8 s. All stimulus blocks were followed by an additional 2 s rest period.

fMRI data acquisition

Scanning was performed on a 3.0 T Siemens MAGNETOM Verio scanner. Researchers[18] scanned 76 interleaved T 2*-weighted axial slices that were 2.0 mm thick without a gap, using a gradient-echo echo multiband echo-planar imaging (MB-EPI) sequence [repetition time (TR) = 2000 ms, echo time (TE) = 43 ms, flip angle (FA) = 80°, field of view (FOV) = 192×192 mm, resolution = 2×2 mm, MB factor = 4]. For anatomical reference, high-resolution T1-weighted images of the whole brain were acquired using a magnetization-prepared rapid acquisition gradient-echo sequence [MPRAGE, TR = 2250 ms, TE = 3.06 ms, inversion time (TI) = 900 ms, FA = 9°, FOV = 256×256 mm, voxel size = 1×1×1 mm].

fMRI data preprocessing

Blood oxygen level-dependent (BOLD) reference images in each run were acquired using a custom methodology of fMRIPrep.[50] Using the BOLD reference, data were motion corrected and slice time corrected, then co-registered to the corresponding T1-weighted images. Then, researchers resampled the registered BOLD time series to the original space.

To create samples, the preprocessed fMRI data were regressed out nuisance parameters, temporally shifted by 4 s for the sake of compensating for hemodynamic delays, despiked to reduce extreme values and averaged within each stimulus block (including the video presentation and rest). Then the data were z-scored across stimulus blocks. This pipeline resulted in a total of 2196 samples, one for each video.

Flatmap construction

We utilized flattened cortical surfaces reconstructed from the anatomical images of individual subjects to visualize the analytical results of the whole cortical region. FreeSurfer[51] was used to generate the cortical surface meshes from T1-weighted anatomical images, followed by five relaxation cuts onto the surface of each hemisphere and removal of the corpus callosum. Finally, functional images were aligned to the anatomical images and projected onto the surface for visualization using Pycortex.[52]

Eliminating spurious correlations between emotion ratings and other related information

To obtain unmixed responses to emotion category ratings, spurious information from visual and semantic responses should be explained away. To this end, we should extract visual and semantic features from video stimuli.

For visual features, motion energy,[53] which is a kind of low-level visual feature, was extracted according to the following steps. First, video frames were spatially downsampled to 96×96 pixels. Then, the images were converted from RGB to (CIE) LAB color space, and the color information was further removed. We employed the output of 6555 motion energy filters, and each filter consisted of a quadrature pair of space-time Gabor filters, which were tuned to six spatial frequencies (0, 2.0, 4.0, 8.0, 16.0, 32.0), three temporal frequencies (0, 2.0, 4.0) and eight spatial direction parameters (0,45.0, 90.0, 135.0, 180.0, 225.0, 270.0, 315.0). Then, the motion energy signals were obtained from the filter output and log-transformed. As a result, 6555 visual features were calculated, and they expressed the preferences for spatial frequencies, temporal frequencies and orientations. The python package used to extract motion energy features from video can be found at https://github.com/gallantlab/pymoten.

For semantic features, semantic ratings for the video stimuli in terms of 73 relatively concrete semantic concepts were provided in an earlier study.[18]

Voxel-wise encoding model fitting and testing

We used a voxel-wise encoding model[20] to predict brain activity evoked by the video stimuli from the corresponding emotion category ratings. Given the necessity of removing spurious correlations between emotion ratings and other related (visual and semantic) information, we performed banded ridge regression[22] to train the encoding model with emotion ratings, motion energy features and semantic features.

Specifically, the brain activity $\mathbf{Y} \in \mathbb{R}^{N \times V}$ was modeled by multiplying the emotion category ratings matrix $\mathbf{X}_1 \in \mathbb{R}^{N \times C_1}$, the motion energy features $\mathbf{X}_2 \in \mathbb{R}^{N \times C_2}$ and the semantic features $\mathbf{X}_3 \in \mathbb{R}^{N \times C_3}$ with the weight matrices $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$, respectively (N = # of samples, V = # of voxels, C_1 = # of emotion categories, C_2 = # of motion energy features, C_3 = # of semantic features). The objective of the banded ridge regression can be formulated as

$$\min \mathcal{L}(\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3) = \| \mathbf{Y} - [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_3] \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \mathbf{W}_3 \end{bmatrix} \|_2^2 + \lambda_1 \| \mathbf{W}_1 \|_2^2 + \lambda_2 \| \mathbf{W}_2 \|_2^2 + \lambda_3 \| \mathbf{W}_3 \|_2^2, \quad \text{(Equation 1)}$$

where the trade-off parameters λ_1, λ_2 and λ_3 control the degree of regularization.

Among the total 61 runs, we selected 51 runs as the training dataset and 10 runs as the test dataset, leading to 1836 training samples and 360 test samples, which was consistent with the setting in a previous study.[18] The optimal regularization parameter was assessed in 10-fold cross-validation using 10 randomly generated training-validation subsets from the training dataset, with 20 different regularization parameters ranging from 10^{-2} to 10^{20}. In the model testing phase, Pearson's correlation coefficients between the predicted brain activity and the measured brain activity in the test dataset were calculated for each voxel individually.

Principal component analysis

We applied PCA to construct the fundamental affective space. First, we selected voxels that were significantly predicted by the model ($p < 0.001$, FDR corrected for multiple comparisons; the number of significant voxels ranged from 9182 to 26126 for each subject). Then, group weights were acquired by vertically concatenating the emotion encoding model weights of selected voxels across the five subjects. PCA was performed on the group weights that yielded the group affective space. To demonstrate the structure of the group affective space, 34 emotion categories were projected onto the two-dimensional space using the loading of PC1 (1st PC) and PC2 as the x-axis and y-axis, respectively. The emotion categories were subsequently colored in red, green and blue according to the relative PCA loadings in PC1, PC2 and PC3, respectively. To understand the cortical organization of the affective space for each subject, we first extracted and normalized the PCA scores from each subject's voxels. Then each cortical voxel was colored according to the PCA scores of PC1, PC2 and PC3, corresponding with the color of the emotion categories in the two-dimensional space. Consequently, we were able to visualize the emotion distribution on the cortical surface.

Cross-subject consistency of the fundamental affective space

To verify the cross-subject consistency of the recovered fundamental affective space, we employed the leave-one-subject-out strategy. Specifically, we constructed a four-dimensional affective space for each subject (serve as individual space) and used the combined data from the remaining four subjects (serve as group space for the excluded subject). For a control space, we further constructed a four-dimensional space by performing PCA on the emotion category ratings of the training dataset directly. To illustrate the consistency of individual spaces, we compared the similarity of the individual space to the group space and that to the control space for each subject using RSA.[31] For each space, we calculated 34×34 RDM based on Pearson's correlation coefficient. Then the similarity between each space pair was measured as the correlation distance of their RDMs.

Comparison between recovered and hypothesized affective dimensions

We used 14 hypothesized affective dimensions that were defined in an earlier study.[21] To compare the estimated affective dimensions to the hypothesized ones, we first defined each hypothesized affective dimension as a 34-dimensional vector, whose elements represent the weights on the emotion categories. According to a previous study,[36] we employed L2-regularized linear regression to obtain the relationship between emotion category ratings and affective dimension ratings. Specifically, we used affective dimensions as predictors to predict each emotion category, which yielded a 14×34 weight matrix. Then, we calculated the correlation between each group PC vector and hypothesized affective dimension vector using Pearson's correlation coefficients.

Smoothness of cortical maps under estimated affective space

To verify that the smoothness of the cortical map is a specific property of the estimated affective space, we compared the cortical map of four-dimensional group affective space with that of any four-dimensional projection of model weights, so that we can rule out the possibility that the model weights themselves smoothly projected onto the cortical sheet. Specifically, we constructed random orthogonal four-dimensional projections by performing singular value decomposition (SVD) on randomly generated 4×34 matrices. These spaces were uniform random rotations of the group affective space. Then, we compared the smoothness of cortical maps between these two spaces.

In order to quantify the defined smoothness, we performed searchlight analysis on voxels using for constructing affective space ($p < 0.001$, FDR-corrected). Firstly, we computed the Euclidean distance between each pair of voxels and rounded up. Although this distance metric does not directly reflect the physical distance due to the presence of gyri in the brain, this has effect on all models that we consider and therefore will not bias the results of our analysis. Then, we projected the encoding model weights of each subject onto the group affective space and random space, respectively. After this, we computed the correlation between the projected weighs for each pair of voxels chosen above.

Finally, for each distance up to eight voxels, we computed the mean correlation between all pairs of voxels separated by that distance. This procedure yields a spatial autocorrelation function for each subject.

QUANTIFICATION AND STATISTICAL ANALYSIS

Given that we need to select the emotion-related voxels to construct the group affective space, we conducted significance test for the voxel-wise encoding model. Specifically, considering that negative correlation is meaningless in the encoding model, one-sided statistical significance was performed by setting the null distribution as correlations between two random vectors from independent Gaussian distribution of the same length as the test dataset. The statistical threshold was calculated by setting $p < 0.01$ for displaying cortical maps of prediction accuracy and $p < 0.001$ for constructing affective space, and multiple comparisons were performed using the FDR procedure.[54]

As for the cross-subject consistency of the affective space, to explore whether the similarity between individual and group space is significantly higher than that between individual and control space, we bootstrapped the model weights for individual and group space, and emotion ratings for control space, by sampling with replacement from the voxel population and the stimuli 1000 times, respectively. The null hypothesis is that the similarity between individual and group space is equal to that between individual and control space. We rejected the null hypothesis if the similarity between individual and control space was never higher than that between individual and group space across the 1000 bootstrap samples (corresponding to $p < 0.001$).

For the smoothness of cortical maps, to estimate the null distribution of smoothness values and obtain statistical significance, we constructed 100 random spaces and regarded the observed mean pairwise correlation under the group affective space as significant if it exceeded all of the 100 random samples which means $p < 0.01$.