PNAS
www.pnas.org

Supplementary Information for

Tracking the affective state of unseen persons

Zhimin Chen & David Whitney

Zhimin Chen
Email: chenzhimin@berkeley.edu

This PDF file includes:

Supplementary text
Figs. S1 to S6
References for SI reference citations

Supplementary Information Text

Participants: In total, we tested 393 healthy participants with normal or corrected-to-normal visual acuity. Sixty-five participants (19 males, 46 females; mean age: 21.5; age range: 18 – 38 years) took part in Experiment 1 in a lab environment. Participants were assigned to different conditions randomly, resulting in 32 participants for the fully-informed condition and 33 for the inferred condition. Participants in Experiment 2, 3a and 3b were tested online through a website that we developed. For Experiment 2, two-hundred and three participants (74 males, 129 females, 2 others; mean age: 21.5; age range: 18 – 37 years) were assigned to 3 different conditions randomly: 51 participants rated clips only in the fully-informed condition; 50 participants rated clips only in the character-only condition; 102 participants rated half of the video clips in the context-only condition and the other half of the videos in the blur-only condition (on average 51 participants for each condition). For Experiment 3a and 3b, seventy-five participants were assigned to three conditions randomly (25 each for the fully-informed condition, the character-only condition and the context-only condition).

Stimuli: All 47 video clips used in our experiments were randomly gathered from online video-sharing website (YouTube) based on the following criteria: 1) showing live action but not animation or monologue; 2) the emotions/affect of the characters should vary across time. These videos portrayed a wide range of social situations (e.g. roadway interactions, interviews, farewells, competition, weddings, etc). There was no data collected on the videos prior to selection. The sets of video clips used in different experiments do not overlap. In Experiment 1, we used 2,844 seconds of 13 video clips from various Hollywood movies that show two main characters interacting with each other had a resolution of 1920 x 1080 and frame rate of 29.97 frames per second. Another 1,214 seconds of 12 video clips from various Hollywood movies were used in Experiment 2. These videos could show two or more than two main characters interacting. In Experiment 3a, we used a total 613 seconds of 9 video clips (8 from Hollywood movies and 1 from a documentary). They all show a single target character alone with no interpersonal interaction. In Experiment 3b, we used a total of 922 seconds of 13 video clips that are non-Hollywood movies, either from home videos or documentaries.

Participants reported how familiar they were with each video clip by selecting a point on a 0-10 continuous scale. The video clips were relatively novel to our participants: the mean self-report familiarity measure across all trials and all participants was quite low (mean: 1.4, SD: 1.4, Range: 0.15 – 5.9). Whether participants had seen the movie or not and how familiar they were with each video clip did not affect the amount of unique variance explained by context ($p > 0.05$ by splitting the data into two groups by familiarity and testing the difference between groups using a permutation test).

To mask out a chosen character, we used video editing software (Adobe Premiere Pro CC) to apply a Gaussian-blurred mask on the face and body of that character frame by frame. We feathered the mask edges to create a contrast modulated envelope, which seamlessly transitions the mask boundary into the video background. The mask was highly blurred so that every detail of the target character was completely invisible. In Experiment 1, all video clips depicted two characters interacting with each other and either of the two characters could be the target and therefore one or the other was masked out. For the videos in Experiment 2, 3a and 3b, one character was chosen as the target and masked out to create video clips for the context-only condition. These masks were then inverted to mask out all contextual background leaving only the target character visible to create video clips for the character-only condition.

In Experiment 1, the processed video clips were presented at full size on a 15-inch Macbook Pro monitor running Matlab and Psychtoolbox *(1, 2)*. Participants sat in a darkened psychophysical experimental booth with a viewing distance of 40 cm. The monitor had a resolution of 1440 x 900 and 60 Hz refresh rate. In Experiment 2, 3a and 3b, the processed video clips were presented within a custom website using an embedded YouTube player. Participants completed the experiment in a non-lab environment. All videos were preloaded prior to the trial to ensure smooth playback.

Methods: Participants first viewed a printed version of the valence-arousal affect rating grid with valence and arousal dimensions depicted. Example words were shown at different locations on the grid, according to the ratings provided by Bradley and Lang (1999). Observers were instructed to familiarize themselves with the dimensions and example word locations before proceeding to the next step. Text indicating the direction of the valence and arousal dimensions were placed on the grid any time during the experiment. Participants were instructed to track and rate the affect of a visible or invisible target character and move a mouse pointer continuously in real-time to different locations inside the affect rating grid to represent the affect they thought the target character was experiencing. Although we used naturalistic videos, observers in our experiment setting were observing people interacting with others or with the environment. All participants completed at least one 2-min practice trial before starting the main experiments. The edited video clips were presented in a random order. Before starting a trial, an example frame from the video with the face and body of the target character was shown on screen and participants were told to track and rate the affect of the specified character but not any other character. Each trial lasted from 58 seconds to 178 seconds. At the end of each trial, participants were told the name and year of the movie from which the video clip was chosen, and were asked to report whether they had seen that movie before. They were also asked to report how familiar they were with the movie and the video clip itself by selecting a point on a 0-10 continuous scale, and how much they liked the video clip. During the experiment, we assessed whether subjects were non-responsive (potentially due to lapsing or other reasons) by calculating the longest duration that the participant kept the mouse pointer in any single location. If the duration was longer than 10 seconds, the participant was reminded to pay more attention in future trials. In all other trials, the participant was given positive feedback. Procedures for Experiment 2 were similar to those in Experiment 1, except that self-reported familiarity and likeness scores were collected using a 0-10 discrete likert scale instead of a continuous one.

Data preprocessing. For all analyses, the continuous rating data were binned into intervals of 100 ms (10 Hz). To ensure the quality of the continuous data, we calculated the longest duration each observer had fixed the mouse pointer in a single stationary location in every trial, and we excluded trials where that duration exceeded 2 standard deviations. This excluded trials where the mouse was physically stationary for greater than ~20 sec, and resulted in the removal of 6.8% of the data. The exclusion of trials in the analysis did not change the significance of the effects reported. We standardized ratings within each participant by subtracting the mean and dividing by the standard deviation of each participant's ratings.

Test-retest reliability of continuous ratings. To test whether our continuous rating method provides consistent and stable ratings from one test administration to the next one, we asked a separate group of 50 participants to rate the same clip twice. The second rating was approximately 1 hour after the first rating of the clip in the fully-informed condition in Experiment 2. The mean Fisher z transformed Pearson correlation coefficients between the initial ratings and the repeated ratings were 0.65 for valence (bootstrapped 95% CI: 0.62 - 0.67) and 0.56 for arousal (bootstrapped 95% CI: 0.53 - 0.59). These second-time ratings were not used in any other analysis in this manuscript. These results confirm that our continuous rating method was robust across multiple administrations.

Split-half correlation. Besides single-subject Pearson correlation, we also used split-half correlation to assess the agreement between subjects rating the same video clip. In split-half correlation, inferred affect ratings provided by different subjects on each video clip were split into two halves, and the averages obtained from ratings by half of the participants were correlated with the averages obtained from ratings by the other half. Across Experiments 1 and 2, we found high between-subject agreement in the inferred affect ratings of the invisible character (see Fig. S2).

Permutation test. Because our method used continuous affect ratings collected from viewing dynamic videos, there is inevitable temporal dependency in the ratings. To evaluate statistical significance, we chose not to use parametric tests (ANOVA and t-test) because they make certain assumptions about the data and its distribution, which would often not hold true for our continuous rating data. Furthermore, we opted not to use time series methods such as ARIMA to assess similarity of ratings between conditions because these techniques require stationarity of the data. Our data was not stationary as examined by Kwiatkowski–Phillips–Schmidt–Shin test of stationarity (3) and could not be transformed to be stationary. Given the aforementioned reasons, we decided to use non-parametric resampling and Monte Carlo permutation methods to generate null distributions of various statistics used in our study; for example, we shuffled the trial labels of whole continuous ratings while preserving the temporal structure in each continuous sequence of ratings. This permutation method preserves all of the temporal structure (dependency or non-stationarity) inherent to continuous ratings but not any video clip-specific information.

We used Monte Carlo permutation tests to evaluate the statistical significance of single-subject correlations (between-subject agreement), split-half correlations (between-subject agreement), Pearson correlations between mean inferred ratings and mean fully-informed ratings (IAT accuracy), and partial correlations between mean inferred ratings and mean fully-informed ratings of the target character when controlling for mean fully-informed ratings of the partner character. The movie clips used in the present study were of various lengths, ranging from 35 seconds to 3 minutes, which could cause problems when averaging or calculating statistics based on continuous ratings from different clips. To deal with this problem, we divided all continuous ratings into 30-seconds data chunks corresponding to different clip periods and calculated statistics for each data chunk separately in permutation tests.

To examine the statistical significance of single-subject correlation, we first calculated the empirical single-subject correlation values by calculating the pairwise correlation coefficient between pairs of affect ratings from different subjects judging the same clip. These correlation values were then averaged across clip periods to obtain an empirical single-subject correlation value. The null distributions were generated by shuffling the video clip labels of continuous ratings within each participant and then recalculating the mean single-subject pairwise correlation coefficient across all pairs of affect ratings from different participants as described above. All averaged correlations were computed by first applying Fisher Z-transformation on all individual correlations, averaging the transformed values, and then transforming the mean back to Pearson's r. Two-tailed p values were calculated by computing the proportion of permuted mean single-subject Pearson correlation coefficients in the null distributions with an absolute value larger than or equal to the absolute value of the empirical mean single-subject Pearson correlation coefficient.

To examine the statistical significance of split-half correlation, we first randomly split ratings by different participants for each clip period into two halves and then calculated correlation coefficients between averaged ratings of the two halves. This process was repeated 1000 times for each clip period and the resulting 1000 split-half correlation coefficients were Fisher z transformed and averaged for each clip period. We then averaged all split-half correlation coefficients across clip periods to obtain an empirical split-half correlation value. The

null distributions were generated by shuffling the video clip labels of continuous ratings provided by each participant for each clip period and then recalculating mean split-half Pearson correlation coefficient across all clip periods as described above. The procedures of averaging across correlations and calculating p values are similar to those used in single-subject correlation described above.

The Monte Carlo permutation and significance testing for Pearson correlation and partial correlation was similar to those used in single-subject correlation described earlier: we shuffled the clip labels of the mean ratings averaged across all participants for each clip. Therefore, the permuted Pearson correlations were calculated between the mean fully-informed ratings and the mean inferred ratings from random clips. The permuted partial correlations were calculated between the inferred ratings and the fully-informed ratings from random clips, while controlling for the fully-informed ratings of the partner character within the same clip as the inferred ratings.

Linear regression analysis. In Experiment 2, 3a and 3b, we used linear regression models to estimate the proportion of unique variance explained only by the context. The full model was constructed by using both the character-only affect ratings and the context-only affect ratings to predict the fully-informed affect ratings of the visible target. A second character-based model was created by using only the character-only ratings to predict the fully-informed ratings of the target. The proportion of unique variance explained only by the context was calculated by subtracting the variance explained by the character-based model from the total amount of variance explained by the full model.

To test whether a non-linear model fit our data better, we compared three regression models with increasing non-linearity and quantified how well the models fitted the data. Model 1 is a linear model; Model 2 is Model 1 added with an interaction term between context-only and character-only variables. Model 3 is Model 2 added with 2-degree polynomial terms of context-only and character-only variables. We quantified how well the models fit the data by calculating the adjusted R-squared, which control for the number of parameters (the degree of freedom in the model). Adding the interaction terms increased the mean adjusted R-squared from 92% (Model 1) to 93.4% (Model 2) and adding the polynomial terms increased it to 94.3% (Model 3). These are very modest improvements; adding non-linear terms did not significantly increase the goodness of fit ($p > 0.5$), and most of the variance has already been explained by a simple linear model.

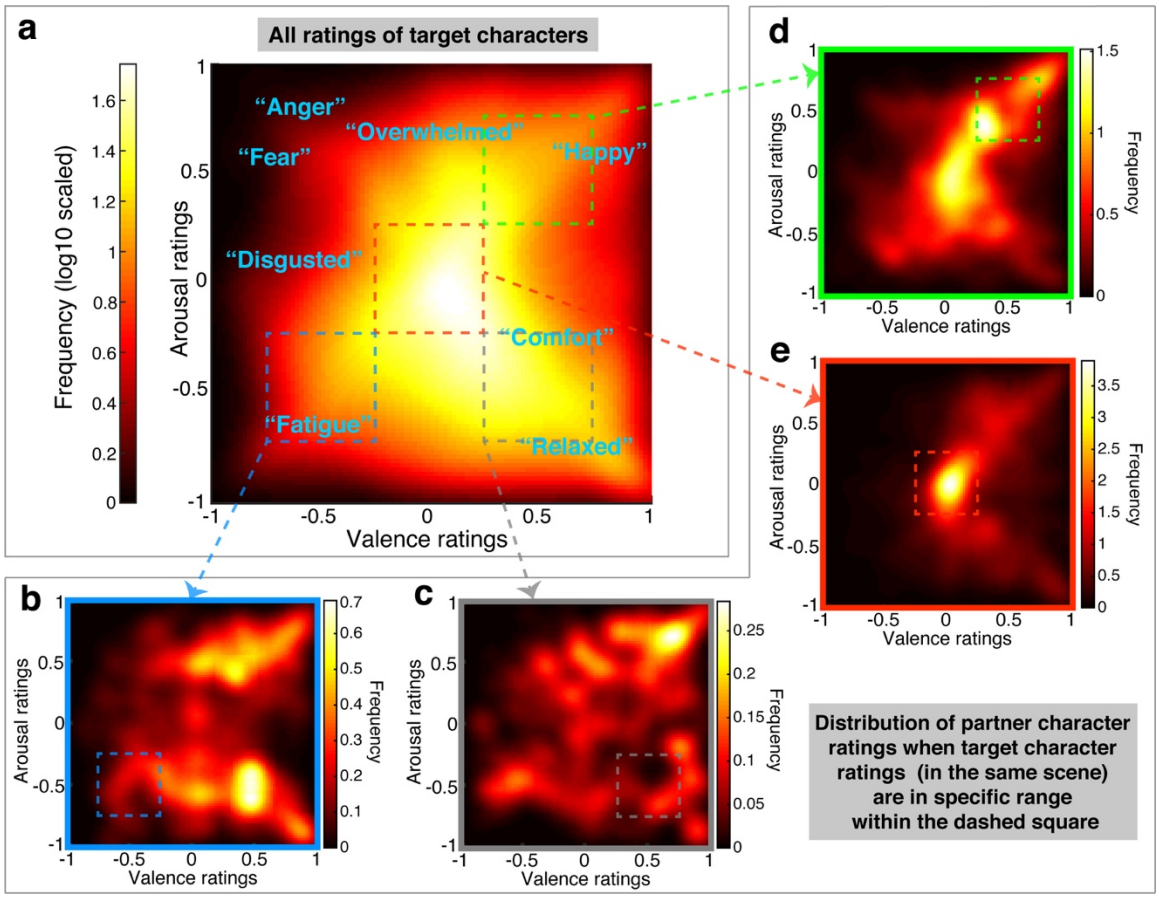

Fig. S1. (a) Distribution of raw fully-informed ratings across the valence-arousal space (data collapsed across Exp 1 and 2). Colors represent the log 10-transformed value of counts within each interval. Affective words have been placed at locations within this space, representing their approximate valence and arousal ratings from Bradley & Lang (1999). (b-e) Qualitative comparison between the distributions of the target characters' affect and those of the corresponding partner characters in the same scene. When the target characters' fully-informed affect ratings are within the range of the dashed rectangle, the distribution of the corresponding partner characters' fully-informed affect ratings were not confined within a certain region or in a certain pattern. The affect of the partner character do not linearly or simply project onto the affect of the target character. They interact in a non-linear, complex way that might be explained in part by contextual information. That the target and partner characters did not always covary in affect helps explain why tracking the partner was not a good proxy for recognizing target character affect (Fig. 2). Ratings were binned into 0.02 intervals and the total number of data points was counted within each 0.02 interval. For visualization, the heatmaps were filtered by a 2-D Gaussian smoothing kernel with a standard deviation of 0.08.

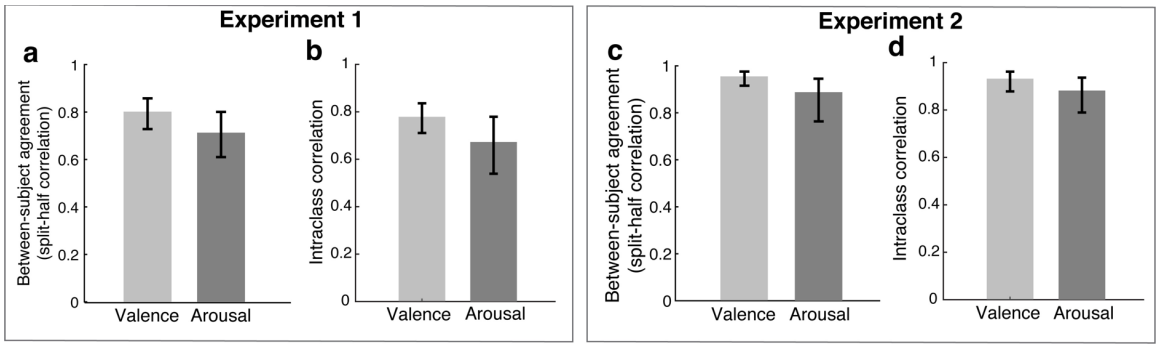

Fig. S2. Between-subject agreement of inferred ratings (when characters were masked and invisible) evaluated by split-half correlation and intraclass correlation. (a&c) In split-half correlation, inferred affect ratings provided by different subjects on each video clip were split into two halves, and the averages obtained from ratings by half of the participants were correlated with the averages obtained from ratings by the other half. (b&d) Intraclass correlation is a common method to assess the conformity of quantitative measurements made by different observers. The total amount of variance in ratings can be divided into variance between subjects and variance between items (time points). Intraclass correlation evaluates between-subject agreement by measuring the proportion of total variance that is between items but not between subjects. Error bars represent bootstrapped 95% confidence interval (CI).

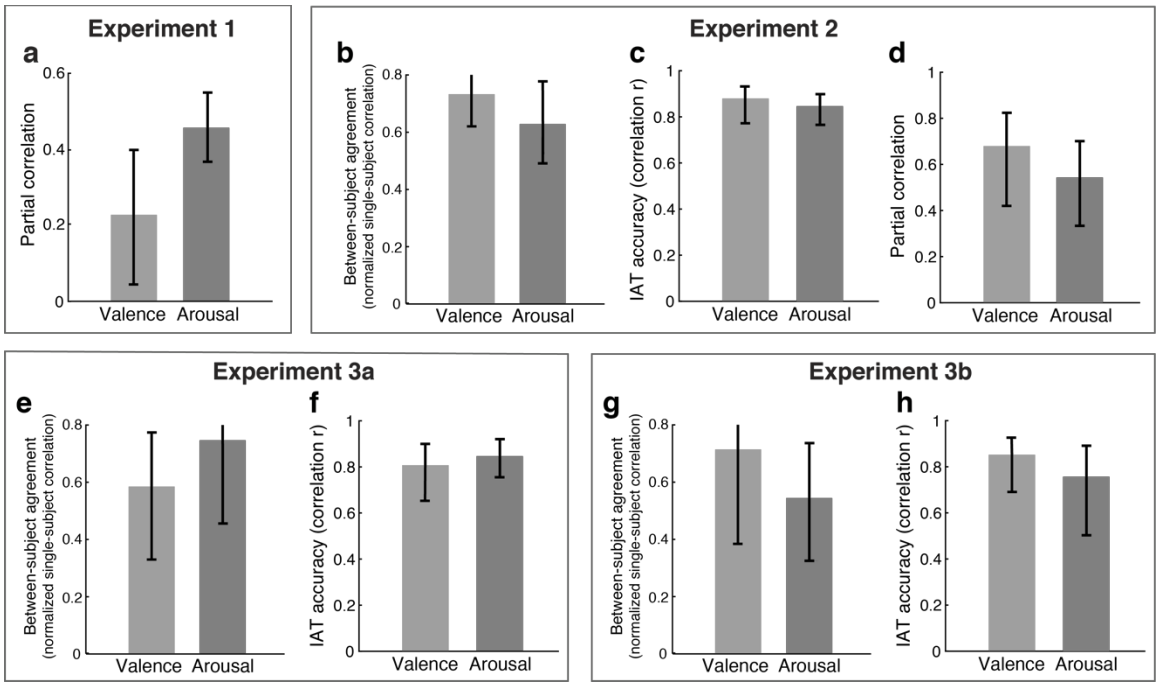

Fig. S3. (a) Mean partial correlations (separately for valence and arousal) between inferred affect ratings of the invisible target and fully-informed affect ratings of the visible target, when controlling for fully-informed affect ratings of the visible partner. (b, e, g) Between-subject agreement of context-only (inferred) affect ratings evaluated by normalized single-subject correlations separately for valence and arousal. (c, f, h) Inferential affective tracking (IAT) accuracy evaluated by mean Pearson correlation coefficients between context-only (inferred) affect ratings of the invisible target character and fully-informed affect ratings of the visible target for valence and arousal. (d) Mean partial correlations between context-only affect ratings and fully-informed affect ratings of the target character, when controlling for the character-only affect ratings of the target character. Error bars represent bootstrapped 95% confidence interval (CI).

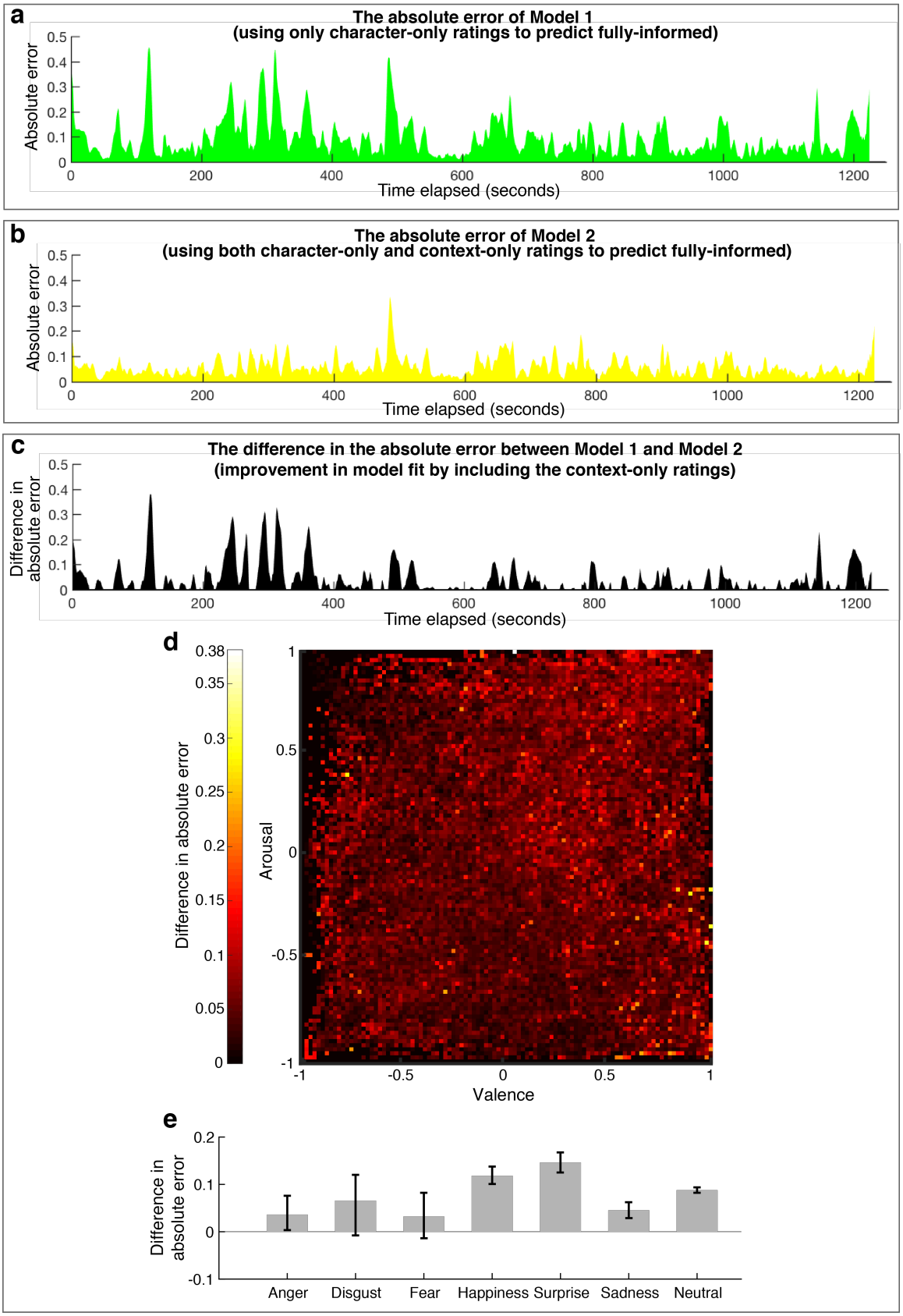
a
The absolute error of Model 1
(using only character-only ratings to predict fully-informed)
Absolute error
Time elapsed (seconds)
b
The absolute error of Model 2
(using both character-only and context-only ratings to predict fully-informed)
Absolute error
Time elapsed (seconds)
c
The difference in the absolute error between Model 1 and Model 2
(improvement in model fit by including the context-only ratings)
Difference in absolute error
Time elapsed (seconds)
d
Difference in absolute error
Arousal
Valence
e
Difference in absolute error
Anger
Disgust
Fear
Happiness
Surprise
Sadness
Neutral

Fig. S4. Contextual information explains significant unique variance in estimating fully-informed affect. (a) The absolute error of Model 1, which uses mean character-only ratings to estimate mean fully-informed affect ratings of the target characters. Essentially, this model tests the role of the face and body of the target character. Videos were concatenated and data were collapsed across dimensions of valence and arousal. (b) The absolute error of Model 2, which uses a linear combination of mean character-only ratings and mean context-only ratings to estimate mean fully-informed affect ratings of the target characters. Essentially, this model tests the role of both context and character specific information in predicting affect ratings. (c) The reduction in prediction error by adding mean context-only ratings to the regression model, calculated by the difference in the absolute error between Model 1 and Model 2. Larger values on the ordinate indicate more improvement with the context. Essentially, this shows how much improvement there is in the model fit by including the context. (d) The benefits of having additional contextual information span the 2D valence and arousal affect space. The color shows the reduction in prediction error by adding additional mean context-only ratings to the regression model, the same as y-axis in (c). Data were binned into 0.2 intervals. (e) The mean benefit of having additional contextual information, the same y-axis as in (c) and (d), in frames containing facial expressions of various emotion categories. We used the Microsoft Azure Emotion API, based on state-of-the art computer vision models *(4)*, to detect frames containing facial expressions of different emotion categories, including anger, contempt, disgust, fear, happiness, neutral, sadness and surprise. We then calculated the mean reduction in prediction error by adding additional mean context-only valence ratings to the regression model across frames labeled as the same emotion category. This analysis was done on data collapsed across Experiment 1 and 2. Error bars represent bootstrapped 95% confidence interval.

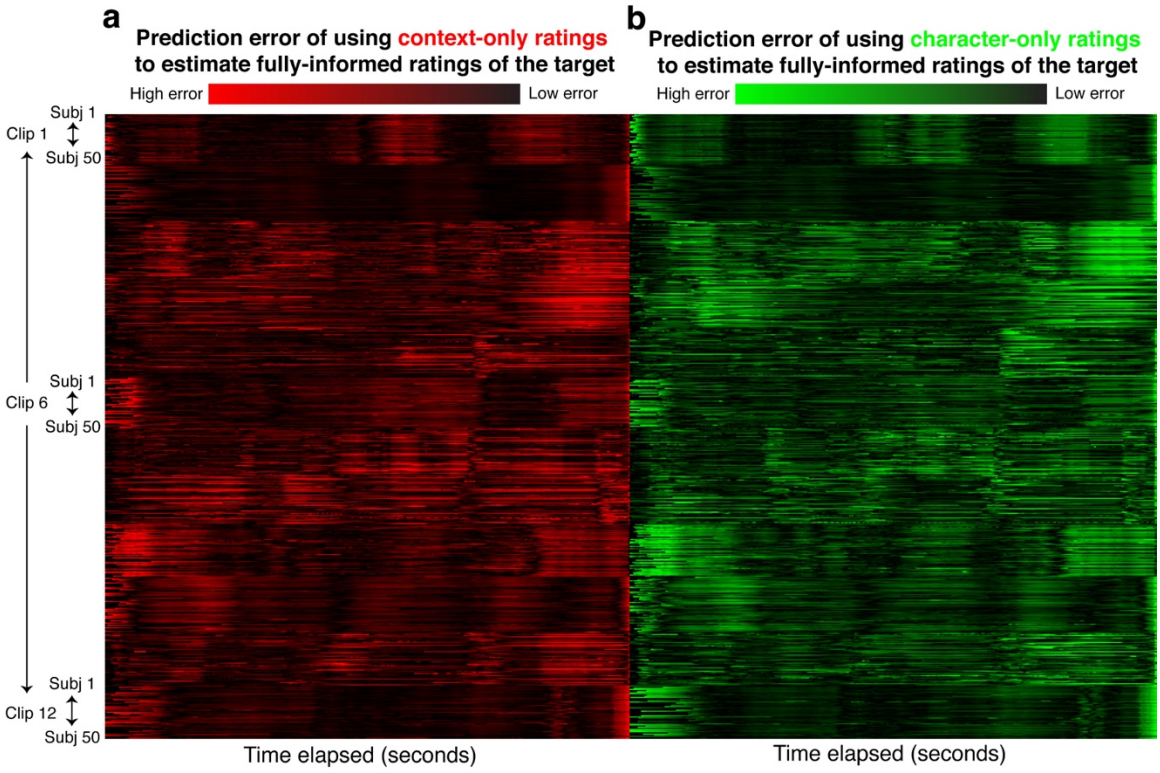

Fig. S5. Affect perceived from only contextual information is nearly as accurate as affect perceived from facial and bodily information. (a) The prediction (absolute) error of using context-only ratings to estimate fully-informed affect ratings of the target characters (indicated by the intensity of the red color; black indicates no model error). Data from different subjects and different videos are stacked together, such that each line indicates a different subject and video. Valence and arousal data were averaged. (b) The prediction (absolute) error of using character-only ratings to estimate fully-informed affect ratings (indicated by the intensity of the green color; black indicates no model error). Note that the intensity of the red color is similar and as broad as the coverage of the green color, showing that the proportion of variance explained by the context is about as high as the variance explained by the face and body. This analysis was done on data collapsed across Experiment 1 and 2.

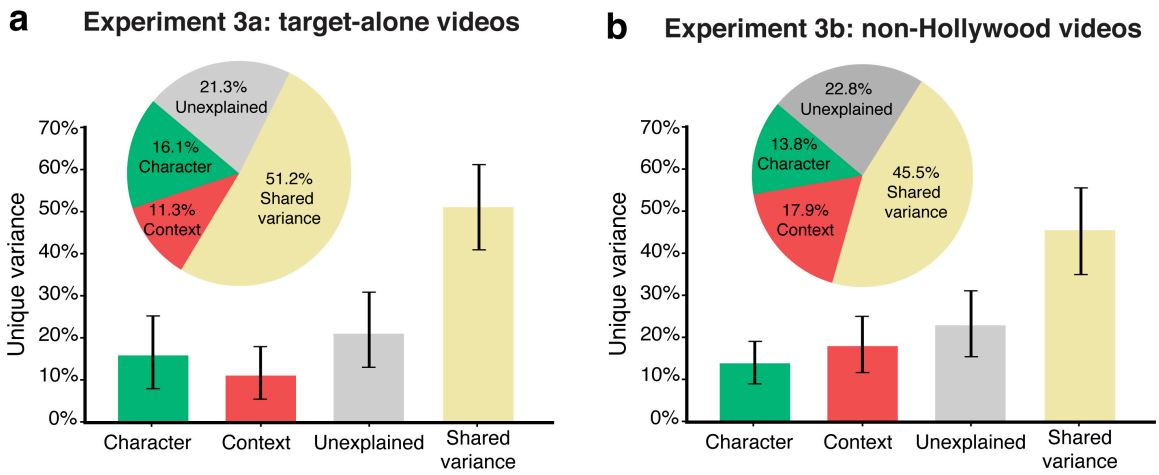

Fig. S6. Proportion of unique variance in the fully-informed affect ratings that could only be explained by context-only affect ratings (in red) versus character-only affect ratings (in green). Yellow bar and pie show the proportion of variance shared between context-only and character-only ratings. (a) Results for Experiment 3a with video clips that show only one target character and no partner character. (b) Results for Experiment 3b with video clips from documentaries and home videos, not Hollywood movies. Error bars represent bootstrapped 95% confidence interval (CI)

References

1. Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10(4):433–436.
2. Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis* 10(4):437–442.
3. Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J Econom 54(1):159–178.
4. Del Sole A (2018) Introducing Microsoft Cognitive Services. *Microsoft Computer Vision APIs Distilled: Getting Started with Cognitive Services*, ed Del Sole A (Apress, Berkeley, CA), pp 1–4.