Tracking the affective state of unseen persons

Zhimin Chen[a,1] and David Whitney[a,b,c]

[a]Department of Psychology, University of California, Berkeley, CA 94720; [b]Vision Science Program, University of California, Berkeley, CA 94720; and [c]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720

Edited by Ralph Adolphs, California Institute of Technology, Pasadena, CA, and accepted by Editorial Board Member Thomas D. Albright February 5, 2019 (received for review July 17, 2018)

Emotion recognition is an essential human ability critical for social functioning. It is widely assumed that identifying facial expression is the key to this, and models of emotion recognition have mainly focused on facial and bodily features in static, unnatural conditions. We developed a method called affective tracking to reveal and quantify the enormous contribution of visual context to affect (valence and arousal) perception. When characters' faces and bodies were masked in silent videos, viewers inferred the affect of the invisible characters successfully and in high agreement based solely on visual context. We further show that the context is not only sufficient but also necessary to accurately perceive human affect over time, as it provides a substantial and unique contribution beyond the information available from face and body. Our method (which we have made publicly available) reveals that emotion recognition is, at its heart, an issue of context as much as it is about faces.

affect | emotion | context | facial expression | visual scene

Emotion recognition is a core human ability, important for understanding others, navigating social environments, and guiding decisions and actions (1, 2). Emotion recognition is also a key component of most measures of so-called emotional intelligence (3), and impairments in emotion recognition are associated with a variety of disorders ranging from autism (4) to schizophrenia (5) to major depression (6).

Emotion recognition is widely assumed to be determined by face and body features, and operational measures of emotion perception or emotional intelligence typically use decontextualized face stimuli (7–11). However, an individual's face and body are usually perceived within a meaningful context, not in isolation. In recent years, there has been growing evidence that perceived emotion in facial expressions is susceptible to contextual influences from several modalities, such as the expresser's tone of voice (11), faces of surrounding people (12), scene gist information (13, 14), and personality traits of the perceiver (15). In the visual domain specifically, recent studies found that emotion recognition from facial expressions is modulated by body posture and the visual scene within which the face is seen (16–21), and this modulation appears to happen routinely and automatically. However, the contribution of context has been difficult to systematically investigate and quantify. Also, the vast majority of experiments used static faces superimposed on disconnected, unrelated, or unnatural visual backgrounds. In contrast, emotion perception is continuous and dynamic in natural environments. As a result, quantifying the role of context in emotion recognition has been elusive, leading authors to treat context as a coarse modulator of perceived emotion, primarily used to disambiguate interpreted facial expressions.

An alternative view is that emotion recognition is, at its heart, a context-based process (21): context makes a significant and direct contribution to the perception of emotion in a precise spatial and temporal manner. Human perceptual systems are exquisitely sensitive to context and gist information in dynamic natural scenes (14, 16, 22–26). Such dynamic gist information could carry rich affect-relevant signals, including the presence of other people, visual background scene information, and social interactions—unique emotional information that cannot be attained from an individual's face and body. For example, a smiling face could accompany completely different internal emotions depending on the context: it could be faked to hide nervousness in an interview setting; it could signal friendliness when celebrating other people's success, and it could also show hostility when teasing or mocking others. Furthermore, much evidence suggests that context is processed rapidly, automatically, and effortlessly when recognizing others' emotions (16, 27–31).

Therefore, we hypothesized that emotion recognition may be efficiently driven by dynamic visual context, independent of information from facial expressions and body postures. We operationalized visual context as the spatial circumstances in which a person is seen. There are other types of context (e.g., stimulus history), but our question focuses on the visual spatial context—all of the visual information available apart from the face and body of the person (e.g., background scene, faces of other people). We investigated whether the visual context alone, in the absence of a person's face and body information, is both sufficient and necessary to recognize the (invisible) person's emotion over time.

To quantify whether dynamic contextual information drives emotion perception, we developed a 3D mouse tracking method to measure an observer's ability to dynamically infer and track emotion in real time: "inferential affective tracking" (IAT) (Fig. 1*A*). It is "inferential" because it explicitly tests the ability to infer the emotional states of other people entirely from contextual cues instead of directly from facial expressions. Similarly, the general method is called "affective tracking" because we measured real-time reporting of affect (valence and arousal) in dynamic videos rather than in static images (*SI Appendix*, *Methods*). Our

Significance

Emotion recognition is widely assumed to be determined by face and body features, and measures of emotion perception typically use unnatural, static, or decontextualized face stimuli. Using our method called affective tracking, we show that observers can infer, recognize, and track over time the affect of an invisible person based solely on visual spatial context. We further show that visual context provides a substantial and unique contribution to the perception of human affect, beyond the information available from face and body. This method reveals that emotion recognition is, at its heart, a context-based process.

Author contributions: Z.C. and D.W. designed research; Z.C. and D.W. performed research; Z.C. contributed new reagents/analytic tools; Z.C. analyzed data; and Z.C. and D.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. R.A. is a guest editor invited by the Editorial Board.

Data deposition: Stimuli and data related to this study have been deposited in the Open Science Framework, https://osf.io/f9rxn/.

See Commentary on page 7169.

[1]To whom correspondence should be addressed. Email: chenzhimin@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1812250116/-/DCSupplemental.

Published online February 27, 2019.

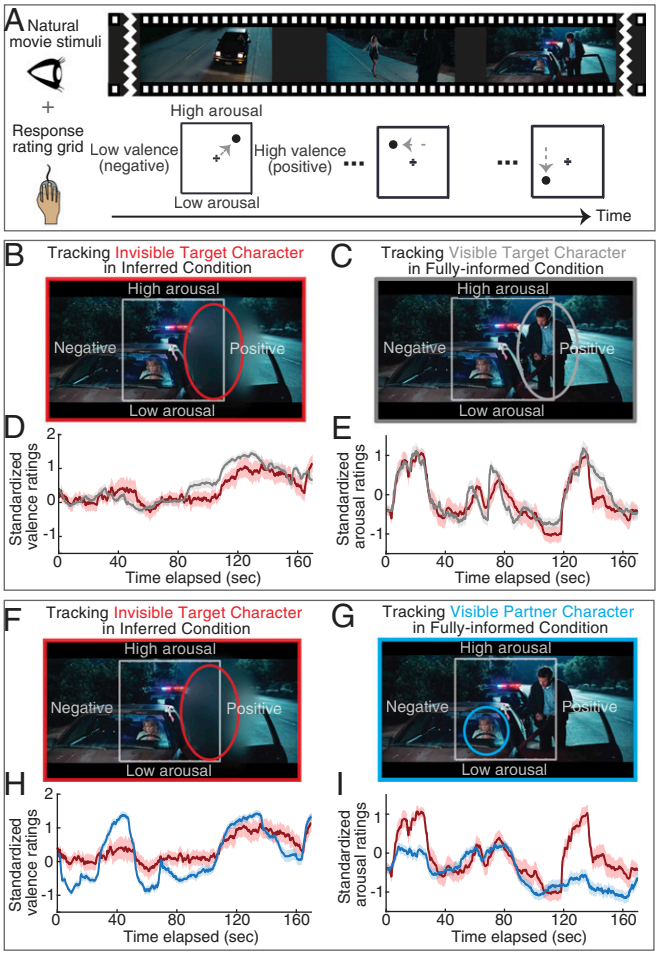

Fig. 1. Experiment 1. (*A*) Observers viewed a silent Hollywood movie clip while moving a mouse pointer within the valence-arousal affect rating grid to continuously report the affect of a chosen character in the video. In the experiments, the affect rating grid was superimposed on top of the video frames. (*B* and *F*) In the inferred condition, the target (the invisible male policeman in this example; circled in red) was occluded by a Gaussian blurred mask, while the partner (the visible female driver) was visible. Participants were asked to infer and track the invisible target's affect. (*C*) In the fully informed condition, participants were asked to track the affect of the target (the male policeman; circled in gray) when everything was visible. (*D* and *E*) Example inferential valence (*D*) and arousal (*E*) ratings over time. Participants' inferred affect ratings of the invisible target (red curve) closely followed the fully informed affect ratings of the visible target (gray curve). (*G*) Participants were asked to track the visible partner (the female driver; circled in blue) in the fully informed condition. (*H* and *I*) Example valence (*H*) and arousal (*I*) ratings. When inferring the affect of the invisible target (red curve), participants did not simply track the affect of the visible partner (blue curve). Shaded regions represent 1 SEM.

experiments focus primarily on valence and arousal because they are the primary dimensions that capture the most variance in affect ratings (32) and the perception of affect has been considered a primitive component of emotion perception (33). We used silent video clips from a variety of sources, including Hollywood movies, home videos, and documentaries totaling 5,593 s across the experiments (*SI Appendix*, *Stimuli*). The video clips included characters interacting with others or with their environment. We removed all auditory and text information to focus on the visual content alone. Movie clips are ideal dynamic stimuli for our experiments because they are widely viewed, they reveal a broad range of emotions (*SI Appendix*, Fig. S1), they are designed specifically to be realistic, and even though they may be staged, they are accepted by audiences. The video clips were gathered from an online video-sharing website, depicting characters in a range of social scenes and emotional situations over a period of time. Videos clips in experiment 1 specifically show two main characters interacting with each other. Experiments 2 and 3 extended this to single or multiple characters and non-Hollywood movie clips. For each video clip, we masked the face and body of a chosen target character frame by frame using a Gaussian blurred mask, such that the target character was completely invisible to viewers (Fig. 1*B*; see *SI Appendix*, *Stimuli*).

In experiment 1, we asked 33 participants to infer and track, in real time, the affect of the invisible target character under what we call the inferred condition (*SI Appendix*, *Methods*). To measure this, we adapted a two-dimensional valence-arousal affect rating grid, previously used to rate static pictures of faces (34). We used a 2D valence-arousal affect-rating grid because it has been shown to be valid and reliable in other domains (35) and it is a uniform space, which allows continuous tracking without predefined categorical boundaries or discontinuities. Moreover, discrete emotional labels do map onto the this 2D space (36), and we confirmed that the distribution of emotions contained in our videos was representative of the full valence-arousal affect space measured using linguistic descriptions (ref. 36; see *SI Appendix*, Fig. S1).

In our experiments, observers moved a mouse pointer within the affect rating grid to continuously report the valence and arousal of an invisible character (Fig. 1*A*). The affect rating grid was superimposed on top of the video, and participants were required to rate the affect of the target continuously in real time while they watched the clip for the first time. Participants were not allowed to view any clip more than once (*SI Appendix*, *Methods*). The affect ratings of target characters were distributed mostly around the center of the affect rating grid with more neutral and medium affect ratings and fewer extreme affect ratings (*SI Appendix*, Fig. S1), showing that the affect in the video clips that we used is not particularly emotionally evocative but is comparable to those in real-world scenarios (36, 37).

Results

Participants agreed with each other about the inferred affect of invisible target characters. We used single-subject Pearson correlation to quantify between-subject agreement. We calculated the pairwise correlation coefficient between pairs of affect ratings from different subjects judging the same clip, which were then divided by single-subject test–retest correlations to obtain normalized values (see *SI Appendix*, Fig. S2 for other measures of between-subject agreement, including split-half correlation and intraclass correlation). These normalized correlation values measure the ratio of the similarity in affect ratings given by different observers relative to the ceiling value, which is the similarity in affect ratings given by the same observer. We found high intersubject agreement in the inferred affect ratings of the invisible character, with a mean normalized single-subject agreement value of 0.61 (bootstrapped 95% CI: 0.50–0.71; $P < 0.001$, permutation tests, see *SI Appendix*) for valence and 0.57 (bootstrapped 95% CI: 0.46–0.68; $P < 0.001$, permutation tests) for arousal (Fig. 2*A*). All mean correlation coefficients 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. Our result indicates that observers agreed with each other about the affect of invisible characters, but it does not yet reveal how accurate they were compared with when the characters were visible.

We measured the accuracy of IAT by comparing inferred affect ratings to affect ratings made when the target character was visible under what we call the "fully informed" condition (Fig. 1*C*). Because there is no absolute ground truth for the expressed affect of the characters on screen, we consider the group consensus of affective

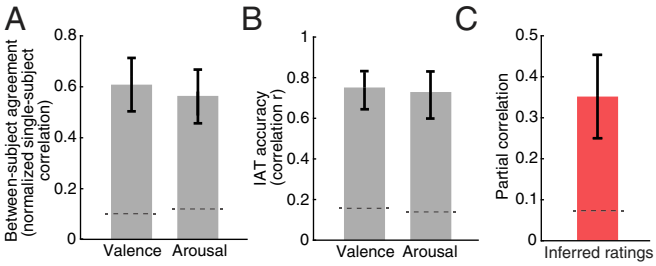

Fig. 2. (*A*) Between-subject agreement evaluated by normalized single-subject Pearson correlation. (*B*) IAT accuracy evaluated by mean Pearson correlation coefficients between inferred affect ratings of the invisible target character and fully informed affect ratings of the visible target. (*C*) Mean partial correlations 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. Error bars represent bootstrapped 95% CI. Dashed lines represent means of permuted null distributions (*SI Appendix, Permutation Test*).

interpretations under the fully informed condition as a practical approximation of ground truth. The fully informed condition includes all of the visual information in the scene, so it is the closest to the default state observers encounter in typical circumstances. To measure the similarity between conditions, we calculated how well the inferred affect ratings (Fig. 1*B*) correlated with the fully informed ratings; we will refer to this measure of similarity as accuracy. To establish the fully informed affect ratings, we asked a different group of 32 participants to track and rate the affect of the target character when he or she was visible on the screen. We chose a between-subject approach to avoid memory effects and interference between conditions. If participants inferred the affect of the invisible target accurately, the inferred affect ratings should closely follow the fully informed affect ratings of the visible target (Fig. 1 *D* and *E*). To quantify IAT accuracy, we computed Pearson correlation coefficients of the time series between inferred affect ratings of the invisible targets and fully informed affect ratings of the same targets when visible. We found a high degree of similarity between inferred affect ratings and fully informed affect ratings, with mean (Fisher *Z* transformed) Pearson correlation coefficients of 0.76 (bootstrapped 95% CI: 0.65–0.83; $P < 0.001$, permutation tests) and 0.73 (bootstrapped 95% CI: 0.60–0.83; $P < 0.001$, permutation tests) for valence and arousal, respectively (Fig. 2*B*). Since between-subject agreement and IAT accuracy were similar for both valence and arousal, we collapsed the data across the two dimensions in the following analyses unless otherwise specified (see *SI Appendix*, Fig. S3 for data pertaining to individual dimensions). In summary, we found that even with no access to any face and body information of the target character, participants were able to accurately infer and track the affect of the invisible target based entirely on contextual cues alone.

One might be concerned that participants simply tracked the affect of the other character who was visibly interacting with the target character (i.e., the partner character) and not actively using dynamic contextual information to infer the affect of the invisible target. To rule out this possibility, we collected affect ratings of the visible partner character in separate trials under fully informed conditions in experiment 1 (no occlusions; Fig. 1*G*). If participants inferred the affect of the invisible target rather than simply tracking the visible partner, we would expect the inferred affect of the invisible target to deviate significantly from the fully informed affect of the visible partner (Fig. 1 *H* and *I*) while still closely following the fully informed affect of the visible target (Fig. 1 *D* and *E*). To quantify this, we calculated partial correlations between inferred and fully informed affect ratings of the target when controlling for fully informed affect ratings of the partner. Separating out the variance attributable to the partner is a conservative approach because characters in an interaction can have covarying affect and emotions (e.g., *SI Appendix*, Fig. S1*E*), and the partner characters should be considered part of the dynamic context rather than just irrelevant information. We found the partial correlation coefficients between inferred affect ratings and fully informed affect ratings of the target character to be strong and significant (mean: 0.35; bootstrapped 95% CI: 0.24–0.45; $P < 0.001$, permutation tests) when accounting for those of the partner (Fig. 2*C* and *SI Appendix*, Fig. S3*A*). This result suggests that when participants were asked to infer and track the invisible target, they did not simply track the visible partner character. The target's affect is more than a linear transformation of the partner's affect: the visual scene background information matters too.

We have shown that the context is sufficient to perceive affect in dynamic and naturalistic environments. However, is the context necessary to most accurately perceive and track affect? Does the context alone possess significant explanatory power beyond the face and body features themselves? To answer these questions, we designed experiment 2 to isolate the contribution of background context information from face and body information. To include a larger variety of scenarios beyond two interacting characters, a new independent set of videos from various Hollywood movies totaling 1,214 s were edited as before (*SI Appendix*, *Stimuli*). Three independent groups of participants were asked to track and rate the affect of a chosen target character in four different conditions: (*i*) fully informed condition, where everything in the clip was visible (Fig. 3*A*); (*ii*) character-only condition, where the context was masked and invisible but the face and body of the target were visible (Fig. 3*C*); (*iii*) context-only condition, where the face and body of the target were masked and invisible but the context was visible (Fig. 3*D*); and (*iv*) blur-only condition, where the target was blurred and the context was replaced by black pixels (Fig. 3*E*). This fourth condition was to

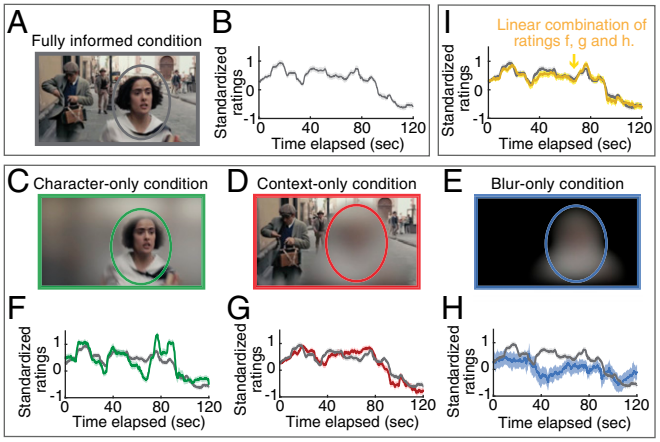

Fig. 3. Experiment 2. (*A*) Fully informed condition: tracking the affect of a visible target in a visible context (the female character in this particular example; circled in gray). (*B*) Example fully informed ratings of the target (gray curve). (*C*) Character-only condition: tracking the visible target (circled in green) while the context was blurred. (*D*) Context-only condition: tracking the blurred target (circled in red) while the context remained visible. (*E*) Blur-only condition: tracking the blurred target (circled in blue) while the context was masked completely by black pixels. (*F*) Example character-only ratings of the target (green curve) compared with fully informed ratings (gray curve). (*G*) Example context-only ratings of the target (red curve) compared with fully informed ratings (gray curve). (*H*) Example blur-only ratings of the blurred target (blue curve) compared with fully informed ratings (gray curve). (*I*) The linear combination of context-only, character-only, and blur-only affect ratings (yellow curve) closely resembled the fully informed rating of the target (gray curve). Shaded regions represent 1 SEM.

control for the residual motion or skin color information available from the blurred target. To show that accurate IAT is not due to individual differences between subjects, one group of participants rated a random half of the video clips in the context-only condition and the other half of the videos in the blur-only condition. On average, video clips in every one of the four conditions were rated by a separate group of 50 participants. We then used linear regression models to measure the degree to which variance in affective tracking is explained only by the character (Fig. 3*D*), the context (Fig. 3*C*), or the blurred mask (Fig. 3*E*).

Similar to experiment 1, participants in experiment 2 accurately inferred the affect of the invisible target character with high agreement. Between-subject agreement evaluated by normalized single-subject correlation was 0.74 (bootstrapped 95% CI: 0.60–0.83; $P < 0.001$, permutation tests) for valence and 0.63 (bootstrapped 95% CI: 0.45–0.77; $P < 0.001$, permutation tests) for arousal (*SI Appendix*, Fig. S3*B*). We also found strong correlations between inferred affect ratings and fully informed affect ratings of the same character, with mean Pearson correlations of 0.88 (bootstrapped 95% CI: 0.77–0.94; $P < 0.001$, permutation tests) and 0.85 (bootstrapped 95% CI: 0.76–0.90; $P < 0.001$, permutation tests) for valence and arousal (*SI Appendix*, Fig. S3*C*).

Is the context necessary to perceive and track affect most accurately, even when face and body information are already available? When controlling for affect ratings in the character-only condition, we found strong and significant partial correlations between affect ratings in the context-only condition and the fully informed condition (mean: 0.61; bootstrapped 95% CI: 0.44–0.73; $P < 0.001$, permutation tests; see Fig. 4*A*). To quantify the size of the unique explanatory power of context, we then used linear regression models to predict mean fully informed affect ratings of the visible target based on mean character-only affect ratings, mean context-only affect ratings, and mean blur-only affect ratings as predictor variables. To account for variance from noise that the regression model could not explain, we normalized the proportion of unique variance by dividing it by the total variance explained by the model. The proportion of unique variance in fully informed affect ratings that could be explained by context-only affect ratings but not character-only affect ratings or blur-only affect ratings was 14.6% (bootstrapped 95% CI: 7.6–22.9%; Fig. 4*B*, red bar) of the total variance explained. Importantly, we found that the benefits of having additional contextual information spanned the whole 2D valence and arousal affect space evenly from neutral to extreme affect ratings (*SI Appendix*, Fig. S4*D*) and across various basic emotion categories annotated by state-of-the-art computer vision models (*SI Appendix*, Fig. S4*E*). Likewise, we also estimated the proportion of unique variance that could only be explained by character-only ratings but not context-only ratings or blur-only ratings (mean: 20.5%; bootstrapped 95% CI: 13.9–28.0%; Fig. 4*B*, green bar), the magnitude of which was comparable to the unique variance explained only by the context ($P > 0.05$, permutation tests). Therefore, the context explains a significant unique portion of variance in the fully informed affect—nearly as much as the character itself (for individual participant data see *SI Appendix*, Fig. S5). In addition, the blur-only ratings contributed only 1.5% of the total variance explained (bootstrapped 95% CI: 0.69–2.5%; Fig. 4*B*, blue bar), which was not different from the permuted null distribution ($P > 0.05$, permutation tests) and was significantly lower than the unique variance of the context ($P < 0.001$, permutation tests). These results suggest that residual information (e.g., kinematics or skin color) in the blurred characters alone was not informative about the affect of characters. While it is conceivable that the contribution of kinematics may be somewhat larger than reported here because the interaction between context and kinematics was not accounted for and might be nonlinear, the key is clearly the presence of context.

We further estimated the proportion of shared variance between character-only and context-only ratings, which reflects the degree of congruency, or the amount of redundant information from target and context. We found that the proportion of shared variance between character-only and context-only ratings was surprisingly high (mean: 58.3%; bootstrapped 95% CI: 53.1–64.4%). This high shared variance suggests that affect recognition is fairly robust, in the sense that one can recognize affect under impoverished conditions, such as when the face, body, or contextual information is missing. Nevertheless, the context does not contain only congruent or redundant information; there is still a significant amount of unique and necessary information available only from the context.

Additional analyses showed that adding nonlinear terms to the model only marginally and nonsignificantly increased the goodness of fit (~1–3% more explained variance), which supports the use of a linear model. Although more complex nonlinear models could, in principle, fit the data better, the linear model provides an excellent fit (89% variance explained) while being parsimonious (see *SI Appendix*, *Linear Regression Analysis*, for a comparison with nonlinear models).

To test whether the contribution of context is essential for scenarios other than Hollywood movie clips or those with interactions between individuals, we conducted experiments 3a and 3b with a new set of video clips. Experiment 3a tested videos that have only one target character and no other character in the scene. Observers could rely on only scene information instead of a partner character's facial expressions to infer the invisible target's affect. Experiment 3b used only nonmovie video clips that were from either documentaries or home videos, rather than from Hollywood movies. One might be concerned that the Hollywood movie clips in experiments 1 and 2 included cinematographer- or director-created emotive environments that might exaggerate the estimated role of the context. However, even if film directors were able to manipulate human affect perception simply with changes to the background scenery, it would support the importance of the context by demonstrating that audiences use this information, which reinforces our point. Artists, including film directors, often reveal the mechanisms and heuristics of visual processing, and this may be another example. Nevertheless, we controlled this in experiment 3b using home videos and documentaries, where the context and facial expressions are not posed or staged in the style of a Hollywood movie. We collected affect ratings from 25 independent observers for

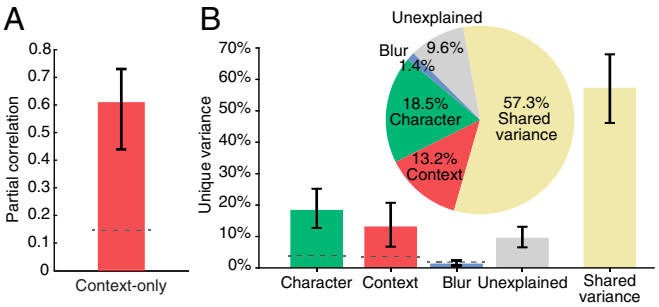

Fig. 4. (*A*) Mean partial correlations between context-only affect ratings and fully informed affect ratings of the target when controlling for the character-only affect ratings of the target. (*B*) Proportion of unique variance in the fully informed affect ratings that could only be explained by context-only affect ratings (in red), character-only affect ratings (in green), and blur-only affect ratings (in blue). Yellow bar and pie show the proportion of variance shared between two or more than two types of ratings. Error bars represent bootstrapped 95% CI. Dashed lines represent means of permuted null distributions (*SI Appendix*, *Permutation Test*).

each of the three conditions: fully informed, context only, and character only (75 new observers in total). The same group of 75 participants in experiment 3a also participated in experiment 3b. We confirmed that participants in experiment 3 accurately inferred the affect of the invisible target character with high agreement. Between-subject agreement evaluated by normalized single-subject correlation was 0.67 (bootstrapped 95% CI: 0.44–0.84; $P < 0.001$, permutation tests) for valence and 0.63 (bootstrapped 95% CI: 0.44–0.79; $P < 0.001$, permutation tests) for arousal (see *SI Appendix*, Fig. S3 *F* and *H* for a breakdown of experiments 3a and 3b). We also found strong correlations between inferred affect ratings and fully informed affect ratings of the same character, with mean Pearson correlations of 0.83 (bootstrapped 95% CI: 0.73–0.90; $P < 0.001$, permutation tests) and 0.80 (bootstrapped 95% CI: 0.67–0.88; $P < 0.001$, permutation tests) for valence and arousal, respectively (*SI Appendix*, Fig. S3 *E* and *G*).

In experiment 3a, which used clips without a partner character, the context contributed a significant amount of unique variance (14.4% of the total explained variance; see *SI Appendix*, Fig. S6*A*), approaching that of the character itself (20.5% of the total explained variance; $P > 0.05$). In experiment 3b, which used more naturalistic videos, the proportion of unique variance explained by the context (23.2% of the total explained variance; see *SI Appendix*, Fig. S6*B*) was even higher than that of the character (17.8% of the total explained variance), although the difference is not statistically significant ($P > 0.05$, permutation tests). These results suggest that visual context is likely to have broad influence on perceived affect across a range of different scenarios.

Discussion

Our results reveal that, in natural scenes, participants use unique information about the background context, independent of any face information, to accurately register and track affect. Even when no face information is present, the visual context is sufficient to infer valence and arousal over time, and different observers agree about the affective information provided by the context. Background contextual information is an essential component of our moment-to-moment emotional experience. It is equally predictive of both neutral and evocative affect (*SI Appendix*, Fig. S4*D*) and different basic emotion categories (*SI Appendix*, Fig. S4*E*). Context is usually taken as having a modulatory influence (14, 19, 20, 27), although recent theories suggest that context might shape and influence the actual perception of emotion signals (21). Our results provide clear evidence that the context by itself is both sufficient and necessary for accurate judgments of the perceived affect of people within that context and contextual information is used even when face and body information is available. The context does not just uniformly amplify or dampen the perceived affect of faces and bodies. Observers actively derive information from contextual information and face and body information and combine them in a context-specific way in real time. Importantly, these substantial contextual influences were observed with a range of different video stimuli, including those with and without interpersonal interactions, with posed or spontaneous facial expressions, and with staged or natural scenes.

What might be the mechanisms underlying such context-based dynamic affect recognition? Numerous empirical findings suggest that human perceptual systems can extract meaningful dynamic gist information from natural scenes efficiently and rapidly (14, 16, 22–26). Such scene gist information could carry emergent properties at a more global or scene-wide scale, which would be accessible through mechanisms of ensemble perception (22). There are a couple of hypotheses about how this information might be used: one hypothesis could be that visual background context is used to support mental simulation of how one would feel in a similar situation, which would be dependent on observers' previous experiences (38). Alternatively, visual context could be integrated in an inferential process based on a set of perceptual or cognitive representations and attributions about other people's mental states (39). Future experiments using our approach with a modified task could be used to distinguish these hypotheses. The more important general point is that context is not at the fringe of emotion recognition, but rather, it may shape and transform emotion into a new holistic interpretation. This might reflect a goal of the visual system: to represent emotion in the most robust way by actively deriving information from context because facial expressions in real life are often absent, ambiguous, or nondiagnostic of emotion (40, 41). In summary, we can better understand the perceptual and neural mechanisms of emotion perception if we incorporate and measure the critical role of contextual information. Our technique allows for this.

Although valence and arousal characterize the dimensional aspect of emotion, they do not fully account for discrete emotion categories such as the difference between anger and fear (33). However, our technique can be extended to categorical emotion as well, and future studies can characterize in detail the conditions or categories under which contextual information might be weighted most strongly. When video frames in our experiment are classified into emotion categories such as happiness, fear, and anger, there is still a significant contribution of context information (*SI Appendix*, Fig. S4*E*), suggesting that our approach can be adopted for use with different emotion spaces (categorical or otherwise).

Our finding suggests that there might be a unique visual mechanism for extracting contextual information to derive affective judgments of people. This has implications for other fields, including the study of emotional intelligence (3) and emotion simulation (42). Although the widely studied construct of emotional intelligence is highly debated (3), most of the major existing emotional intelligence tests include some form of emotion perception, recognition, or emotion understanding measure. These measures usually rely on static, isolated, and decontextualized pictures of faces. Our results suggest that any test of emotional intelligence that incorporates a perceptual measure of emotion recognition or emotion understanding (9, 10) needs to be revised to take into account the separate but nearly as important factor of context in emotion recognition. An individual may be able to recognize static photos of facial emotions but fail to actually understand the displayed emotion unless they successfully take into account the context.

Emotional inference is equally important for computer vision, which is at a stage now where machines are increasingly able to recognize emotion with high accuracy in images and videos based on facial expressions (43). However, our results reveal that human recognition of affect goes well beyond accessing image-level features of a face. Instead, emotion recognition depends strongly on the specific context. As computer vision models of emotion recognition are increasingly incorporated into daily life, such as security surveillance, personalized marketing, and social media, it will be important to understand how humans actually recognize emotion in the real world. Recent efforts to incorporate the context have found that neural networks achieved moderately higher accuracy when both body and contextual information were used as inputs rather than just body alone (44). Although these models are nowhere near as accurate as human observers, the approach of using the context is promising. Indeed, our results demonstrate that recognition of emotion is, at its heart, an issue of context as much as it is about facial and body expressions. Computer vision, neural, and social cognitive models, as well as psychological measures of emotional intelligence, will benefit by taking this into account.

Materials and Methods

We used 47 video clips in total in our experiments. The sets of video clips used in different experiments do not overlap. Our library of videos, affect ratings, and analysis code have been made available at https://osf.io/f9rxn/. In total, we tested 393 healthy participants (205 females, mean age = 22.5 y). All

participants gave informed consent before starting experiments. All participants were naive to the purpose of the experiment. The study was approved by institutional review at the University of California, Berkeley. Participants, stimuli, methods, and analyses are outlined in detail in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Allison Yamanashi Leib, Iris Mauss, Ken Nakayama, and Mauro Manassi for their feedback on earlier drafts of this article. We also thank Cristina Ghirardo, Dylan Medlock, Jiaming Zou, Kevin Huang, Luciano Lopez, and Yitian Ma for assistance with video editing.

1. Keltner D, Haidt J (1999) Social functions of emotions at four levels of analysis. *Cogn Emotion* 13:505–522.
2. Olsson A, Ochsner KN (2008) The role of social cognition in emotion. *Trends Cogn Sci* 12:65–71.
3. Mayer JD, Roberts RD, Barsade SG (2008) Human abilities: Emotional intelligence. *Annu Rev Psychol* 59:507–536.
4. Harms MB, Martin A, Wallace GL (2010) Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychol Rev* 20:290–322.
5. Kohler CG, Walker JB, Martin EA, Healey KM, Moberg PJ (2010) Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophr Bull* 36:1009–1019.
6. Dalili MN, Penton-Voak IS, Harmer CJ, Munafò MR (2015) Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychol Med* 45:1135–1144.
7. Ekman P (1992) An argument for basic emotions. *Cogn Emotion* 6:169–200.
8. Matsumoto D, et al. (2000) A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian brief affect recognition test (JACBART). *J Nonverbal Behav* 24:179–209.
9. Mayer JD, Salovey P, Caruso DR, Sitarenios G (2003) Measuring emotional intelligence with the MSCEIT V2.0. *Emotion* 3:97–105.
10. Russell JA, Dols JMF (1997) *The Psychology of Facial Expression* (Cambridge Univ Press, Cambridge, UK).
11. Paulmann S, Pell MD (2010) Contextual influences of emotional speech prosody on face processing: How much is enough? *Cogn Affect Behav Neurosci* 10:230–242.
12. Masuda T, et al. (2008) Placing the face in context: Cultural differences in the perception of facial emotion. *J Pers Soc Psychol* 94:365–381.
13. de Gelder B, Van den Stock J (2011) Real faces, real emotions: Perceiving facial expressions in naturalistic contexts of voices, bodies and scenes. *The Handbook of Face Perception* (Oxford Univ Press, New York), pp 535–550.
14. Righart R, de Gelder B (2008) Recognition of facial expressions is influenced by emotional scene gist. *Cogn Affect Behav Neurosci* 8:264–272.
15. Calder AJ, Ewbank M, Passamonti L (2011) Personality influences the neural responses to viewing facial expressions of emotion. *Philos Trans R Soc Lond B Biol Sci* 366:1684–1701.
16. Aviezer H, Bentin S, Dudarev V, Hassin RR (2011) The automaticity of emotional face-context integration. *Emotion* 11:1406–1414.
17. Aviezer H, Trope Y, Todorov A (2012) Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338:1225–1229.
18. Barrett LF, Kensinger EA (2010) Context is routinely encoded during emotion perception. *Psychol Sci* 21:595–599.
19. Kayyal M, Widen S, Russell JA (2015) Context is more powerful than we think: Contextual cues override facial cues even for valence. *Emotion* 15:287–291.
20. Wieser MJ, Brosch T (2012) Faces in context: A review and systematization of contextual influences on affective face processing. *Front Psychol* 3:471.
21. Aviezer H, Ensenberg N, Hassin RR (2017) The inherently contextualized nature of facial emotion perception. *Curr Opin Psychol* 17:47–54.
22. Whitney D, Yamanashi Leib A (2018) Ensemble perception. *Annu Rev Psychol* 69:105–129.
23. Leib AY, Kosovicheva A, Whitney D (2016) Fast ensemble representations for abstract visual impressions. *Nat Commun* 7:13186.
24. Righart R, de Gelder B (2008) Rapid influence of emotional scenes on encoding of facial expressions: An ERP study. *Soc Cogn Affect Neurosci* 3:270–278.
25. Kret ME, Roelofs K, Stekelenburg JJ, de Gelder B (2013) Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. *Front Hum Neurosci* 7:810.
26. Mavratzakis A, Herbert C, Walla P (2016) Emotional facial expressions evoke faster orienting responses, but weaker emotional responses at neural and behavioural levels compared to scenes: A simultaneous EEG and facial EMG study. *Neuroimage* 124:931–946.
27. Barrett LF, Mesquita B, Gendron M (2011) Context in emotion perception. *Curr Dir Psychol Sci* 20:286–290.
28. Mumenthaler C, Sander D (2015) Automatic integration of social information in emotion recognition. *J Exp Psychol Gen* 144:392–399.
29. Meeren HKM, van Heijnsbergen CCRJ, de Gelder B (2005) Rapid perceptual integration of facial expression and emotional body language. *Proc Natl Acad Sci USA* 102:16518–16523.
30. Righart R, de Gelder B (2006) Context influences early perceptual analysis of faces—An electrophysiological study. *Cereb Cortex* 16:1249–1257.
31. de Gelder B, et al. (2006) Beyond the face: Exploring rapid influences of context on face processing. *Prog Brain Res* 155:37–48.
32. Feldman LA (1995) Valence focus and arousal focus: Individual differences in the structure of affective experience. *J Pers Soc Psychol* 69:153–166.
33. Russell JA (2003) Core affect and the psychological construction of emotion. *Psychol Rev* 110:145–172.
34. Russell JA, Weiss A, Mendelsohn GA (1989) Affect grid: A single-item scale of pleasure and arousal. *J Pers Soc Psychol* 57:493–502.
35. Lang PJ, Bradley MM, Cuthbert BN (1997) International affective picture system (IAPS): Technical manual and affective ratings (NIMH Center for the Study of Emotion and Attention, University of Florida, Gainesville, FL), pp 39–58.
36. Bradley MM, Lang PJ (1999) Affective norms for English words (ANEW): Instruction manual and affective ratings (NIMH Center for the Study of Emotion and Attention, University of Florida, Gainesville, FL).
37. Kossaifi J, Tzimiropoulos G, Todorovic S, Pantic M (2017) AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis Comput* 65:23–36.
38. Gallese V, Sinigaglia C (2011) What is so special about embodied simulation? *Trends Cogn Sci* 15:512–519.
39. Gopnik A, Wellman HM (1994) The theory theory. *Mapping the Mind: Domain Specificity in Cognition and Culture*, eds Hirchfield LA, Gelman SA (Cambridge Univ Press, New York).
40. Fernández-Dols JM, Crivelli C (2013) Emotion and expression: Naturalistic studies. *Emot Rev* 5:24–29.
41. Wenzler S, Levine S, van Dick R, Oertel-Knöchel V, Aviezer H (2016) Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion* 16:807–814.
42. Zhou H, Majka EA, Epley N (2017) Inferring perspective versus getting perspective: Underestimating the value of being in another person's shoes. *Psychol Sci* 28:482–493.
43. Dhall A, et al. (2017) From individual to group-level emotion recognition: EmotiW 5.0. *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (ACM, New York), pp 524–528.
44. Kosti R, Alvarez JM, Recasens A, Lapedriza A (2017) Emotion recognition in context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York), pp 1667–1675.