Title

The Role of Visual Context in Emotion Recognition

Author

Chen, Zhimin

Publication Date

2020

Peer reviewed|Thesis/dissertation

The Role of Visual Context in Emotion Recognition

By

Zhimin Chen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor David Whitney, Chair
Professor Richard Ivry
Professor Iris Mauss

Fall 2020

Abstract

The Role of Visual Context in Emotion Recognition

by

Zhimin Chen

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor David Whitney, Chair

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 unnatural, static, or decontextualized conditions. However, an individual's face and body are usually perceived within a meaningful context, not in isolation. The visual context, therefore, may provide useful or even necessary information when interpreting emotion. Here, we investigated the role of visual context in dynamic emotion recognition. First, we developed a novel method, "inferential affective tracking (IAT)", to reveal and quantify the contribution of visual context to affect (valence and arousal) perception. We show that 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. Next, we tested the efficiency of IAT by measuring the speed of recognizing emotion from contextual information alone. Using cross-correlation analyses, we found that inferring affect based on visual context alone is just as fast as tracking affect with all available information including face and body. We further demonstrated with empirical evidence that this approach has high precision in detecting a sub-second temporal lag. Finally, we extended and adapted the IAT method to test categorical emotion perception rather than affect. This method is very similar to the IAT technique and so we call it "inferential emotion tracking (IET)". Using IET, we show that the presence of visual context can override interpreted emotion categories from face and body information. Strikingly, we find that visual context determines perceived emotion nearly as much and as often as face and body information does. Taken together, these experiments reveal that emotion recognition is, at its heart, an issue of context as much as it is about faces. Seemingly complex context-based emotion perception is far more efficient than previously assumed.

Table of Contents

Acknowledgements

This work would not have been possible without the support and assistance of many people.

I would first like to thank the members of the Whitney lab over the last five years: Allison Yamanashi Leib, Kathy Zhang, Wesley Chaney, Ye Xia, Zixuan Wang, Susan Hao, Zhihang Ren, Teresa Canas-Bajo, Yuki Murai, and Cristina Ghirardo. It has been a great pleasure to work with such a dedicated group of scientists.

I would like to thank all my scientific mentors. I would first like to thank Gerrit Maus for taking me as a research assistant in the Whitney Lab. I have been fascinated with perception and the brain ever since. I would like to give a huge thank you to David Whitney for taking me as a graduate student. I am very grateful for his guidance and support in helping me grow as a scientist. It is an honor to learn from someone who is so passionate and knowledgeable about perception. I would also like to thank Ken Nakayama for sharing with me his wisdom on science and life. I would also like to thank Bill Prinzmental and Erv Hafter for their valuable feedback on my research. I would also like to thank my thesis committee, Richard Ivry and Iris Mauss, for their great research advice and encouragement.

I would like to thank my family and friends for their support throughout graduate school. I would especially like to thank my husband Jeffrey Zhang. He has been incredibly supportive of everything I do and believing in my ability. My research would not have gone as smoothly without the emotional and technical support from him.

Chapter 1: Introduction

The ability to accurately recognize other people's emotion is critical for social functioning. Emotion recognition is often a core element in widely used measures of emotional intelligence (Mayer et al., 2008). Impairments in emotion recognition are also linked to a range of mental disorders such as autism (Harms et al., 2010) to schizophrenia (Kohler et al., 2010) to major depression (Dalili et al., 2015). Furthermore, understanding emotion recognition is important for building computer vision models to recognize emotion efficiently and accurately, like humans do.

Emotion recognition is widely assumed to be determined primarily by facial expressions and body features, and models of emotion recognition have mainly focused on facial and bodily features in static, unnatural, or de-contextualized conditions. However, faces are usually seen within situational contexts in daily life, and humans can seamlessly integrate contextual information in the process of recognizing emotion. Recent studies found that emotion recognition from facial expressions is modulated by contextual influences such as the body posture and the visual scene within which the face is seen, and this appears to happen routinely and automatically (e.g. Aviezer et al., 2011; Aviezer et al., 2017; Barrett & Kensinger, 2010). Despite these scientific advances, contextual information is still often regarded as secondary to facial information, mainly incorporated to modulate or disambiguate perceived emotion in faces. This might be because most of the past experiments used static face stimuli superimposed on disconnected, unrelated, or unnatural visual backgrounds. In contrast, emotion perception is continuous and dynamic in natural environments.

An alternative hypothesis is that emotion recognition may be substantially driven by context. Visual context can convey unique emotional information that cannot be attained from an individual's face and body, such as the emotion of other people, background scene, and social interactions. The goal of Chapter 2 is to examine this hypothesis. 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 test this, we developed a 3D mouse tracking method, called "inferential affective tracking (IAT)", to measure an observer's ability to dynamically infer and track affect (valence and arousal) in real time while viewing dynamic videos. We examined the role of visual context by masking out the face and body of a chosen character in the video and leaving only the contextual information visible. We predicted that observers can make reasonably accurate predictions of the invisible character's affect when only the contextual information is available. We also introduce a new analysis to directly quantify and compare the amount of unique information provided by context versus face and body features.

Whether visual context is primary and useful to emotion recognition depends on the speed of the available information from the context. If context is as efficient (fast) as when the face and body information are present, it would suggest that context plays a pivotal and primary role, rather than secondary or modulatory one. Previous literature has typically regarded the use of context alone, in the absent of facial expressions, to be indirect and deliberate because we have to rely on abstract causal principles rather than direct perceptual cues (Ekman, 1992; Skerry & Saxe, 2014). Further, conceptual models of perception tend to integrate context or scene effects only at a relatively late and high-level processing stage (Bar, 2004). The implication of these previous studies is that contextual effects on emotion recognition might be relatively slow, or at least slower than the recognition of a facial expression. A competing hypothesis is that

contextual information could be processed with a short latency or in parallel with facial expression information, in which case the latency of inferred emotion perception could be very brief. Chapter 3 examined this by measuring the speed of inferring emotion from contextual information alone (IAT), relative to the speed of recognizing emotion with all available information including facial expression. We further provide additional experiments to support and validate the precision of this approach in measuring the speed of context-based emotion perception in the millisecond (msec) range.

A limitation of the IAT method introduced in Chapter 2 and 3 is that it only characterizes the influence of context on the affective dimensions of valence and arousal. However, this dimensional representation of affect is not the same as discrete emotion categories. For example, distinct emotion categories like anger and fear might have similar values in terms of valence and arousal (Bradley & Lang, 1999; Warriner et al., 2013). It is not a priori obvious that context would or should play a critical role in the perception of discrete emotion categories in dynamic natural scenes. Discrete emotion categories are often considered to be expressed with certain facial features and movements (Cordaro et al., 2018; Ekman, 1992; Matsumoto et al., 2008). Therefore, it remains to be tested whether context is necessary to perceive discrete emotion categories in natural dynamic scenes. In Chapter 4, we extended the IAT technique to test categorical emotion perception rather than affect. This method is very similar to the IAT technique and so we call it "inferential emotion tracking (IET)". We examined whether the visual context is necessary to most accurately perceive emotion categories, even when face and body information is available.

Although there are many prior studies demonstrating that context can influence the perception of facial expressions (Aviezer et al., 2017; Wieser & Brosch, 2012), none of them predict that contextual information alone could be sufficient and necessary to accurately recognize emotion. Viewed in the light of a Bayesian account, emotions are internal experiences and observers need to perform an inverse inference to use observed effects to infer underlying emotions. Facial expression, as a source of signals that support emotion inference, has been suggested to be inherently noisy, ambiguous, and uncertain (Hassin et al., 2013; Russell, 2016). Context may therefore be a primary cue to emotion and carry rich affect-relevant signals that help represent emotion in the most robust way. Our tracking technique allows us to test this by quantitatively characterizing the unique contribution and the processing speed of contextual information in the absence of facial signals with different emotion spaces (categorical or continuous). Together, the three chapters presented here will help us better understand the role of context, which will be beneficial to computer vision, neural, and social cognitive models, as well as psychological measures of emotional intelligence.

Chapter 2: Inferential affective tracking (IAT): tracking the affective state of unseen persons

The research in this chapter has been published in the following article and is included with permission:

Chen, Z., & Whitney, D. (2019). Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, *116*(15), 7559-7564.

Introduction

Emotion recognition is a core human ability, important for understanding others, navigating social environments, and guiding decisions and actions (Keltner & Haidt, 1999; Olsson & Ochsner, 2008). Emotion recognition is also a key component of most measures of so-called emotional intelligence (Mayer et al., 2008), and impairments in emotion recognition are associated with a variety of disorders ranging from autism (Harms et al., 2010) to schizophrenia (Kohler et al., 2010) to major depression (Dalili et al., 2015).

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 (Ekman, 1992; Matsumoto et al., 2000; Mayer et al., 2003; Russell & Dols, 1993). 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 (Paulmann & Pell, 2010), faces of surrounding people (Masuda et al., 2008), scene gist information (de Gelder & Van den Stock, 2011; Righart & de Gelder, 2008a), and personality traits of the perceiver (Calder et al., 2011). 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 (Aviezer et al., 2011; Aviezer et al., 2012; Aviezer et al., 2017; Barrett & Kensinger, 2010; Kayyal et al., 2015; Wieser & Brosch, 2012), 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 (Aviezer et al., 2017): 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 (Aviezer et al., 2011; Kret et al., 2013; Mavratzakis et al., 2016; Righart & de Gelder, 2008a; Righart & de Gelder, 2008b Whitney & Yamanashi Leib, 2016; Yamanashi Leib et al., 2016). 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 (Aviezer et al., 2011; Barrett et al., 2011; de Gelder et al., 2006; Mumenthaler et al., 2015; Meeren et al., 2005; Righart & de Gelder, 2006).

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. 2.1A). 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 (see *Methods* in *Materials and Methods*). Our experiments focus primarily on valence and arousal because they are the primary dimensions that capture the most variance in affect ratings (Feldman, 1995) and the perception of affect has been considered a primitive component of emotion perception (Russell, 2003). We used silent video clips from a variety of sources, including Hollywood movies, home videos, and documentaries totaling 5,593 s across the experiments (see *Stimuli* in *Materials and Methods*). 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 (Fig. A1 in *Appendix A*), 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. 2.1B; see *Stimuli* in *Materials and Methods*).

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 (see *Methods* in *Materials and Methods*). To measure this, we adapted a two-dimensional valence-arousal affect rating grid, previously used to rate static pictures of faces (Russell et al., 1989). We used a 2D valence-arousal affect-rating grid because it has been shown to be valid and reliable in other domains (Lang et al., 1997) 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 (Bradley & Lang, 1999), 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 (Bradley & Lang, 1999; see Fig. A1 in *Appendix A*).

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. 2.1A). 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 (see *Methods* in *Materials and 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 (Fig. A1 in *Appendix A*), 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 (Bradley & Lang, 1999; Kossaifi et al., 2017).

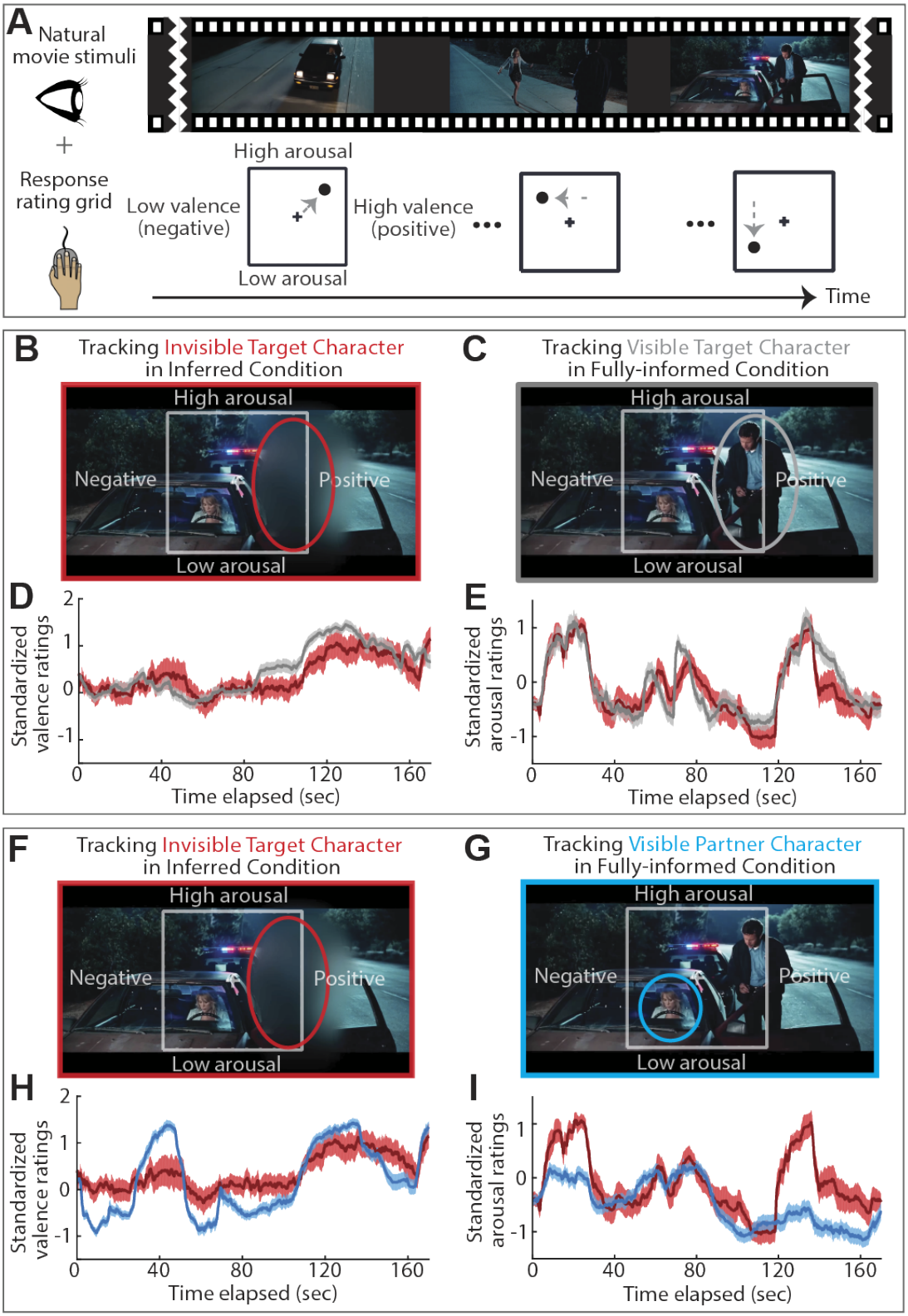


Fig. 2.1. Experiment 1. (A) Observers viewed a silent Hollywood movie clip while moving a mouse pointer within the valence-arousal affect rating grid to continuously report the affect of a chosen character in the video. In the experiments, the affect rating grid was superimposed on top of the video frames. (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.

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 Fig. A2 in *Appendix A* 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 inter-subject 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 *Permutation Test* in *Materials and Methods*) for valence and 0.57 (bootstrapped 95% CI: 0.46–0.68; $p < 0.001$, permutation tests) for arousal (Fig. 2.2A). 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. 2.1C). Because there is no absolute ground truth for the expressed affect of the characters on screen, we consider the group consensus of affective 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. 2.1B) 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. 2.1 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. 2.2B). 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 Fig. A3 in *Appendix A* 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.

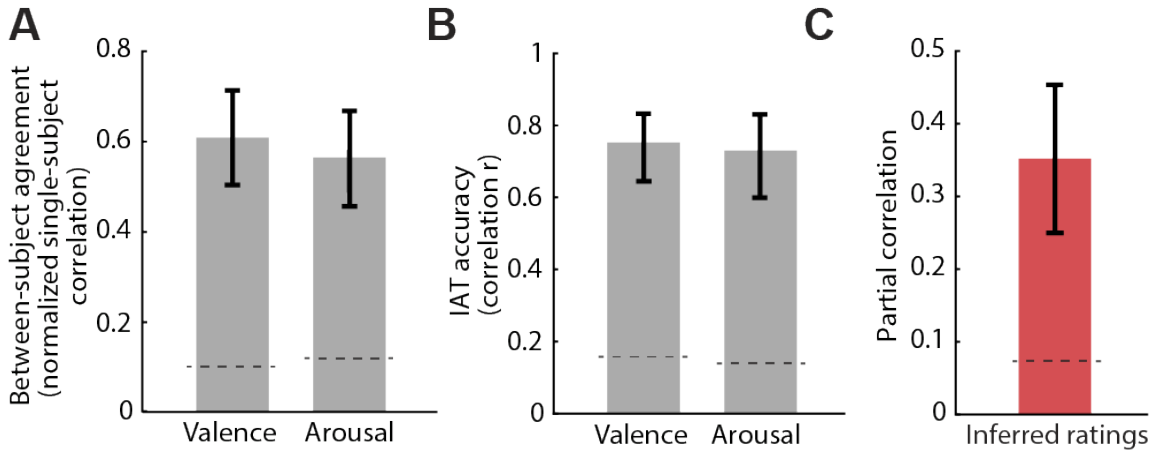


Fig. 2.2. (A) Between-subject agreement evaluated by normalized single-subject Pearson correlation. (B) IAT accuracy evaluated by mean Pearson correlation coefficients between inferred affect ratings of the invisible target character and fully informed affect ratings of the visible target. (C) Mean partial 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 (see *Permutation Test* in *Materials and Methods*).

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. 2.1G). 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. 2.1 H and I) while still closely following the fully informed affect of the visible target (Fig. 2.1 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., Fig. A1E in *Appendix A*), 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. 2.2C and Fig. A3A in *Appendix 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 (see *Stimuli* in *Materials and Methods*). 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. 2.3A); (ii) character-only condition, where the context was masked and invisible but the face and body of the target were visible (Fig. 2.3C); (iii) context-only condition, where the face and body of the target were masked and invisible but the context was visible (Fig. 2.3D); and (iv) blur-only condition, where the target was blurred and the context was replaced by black pixels (Fig. 2.3E). This fourth condition was to 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. 2.3D), the context (Fig. 2.3C), or the blurred mask (Fig. 2.3E).

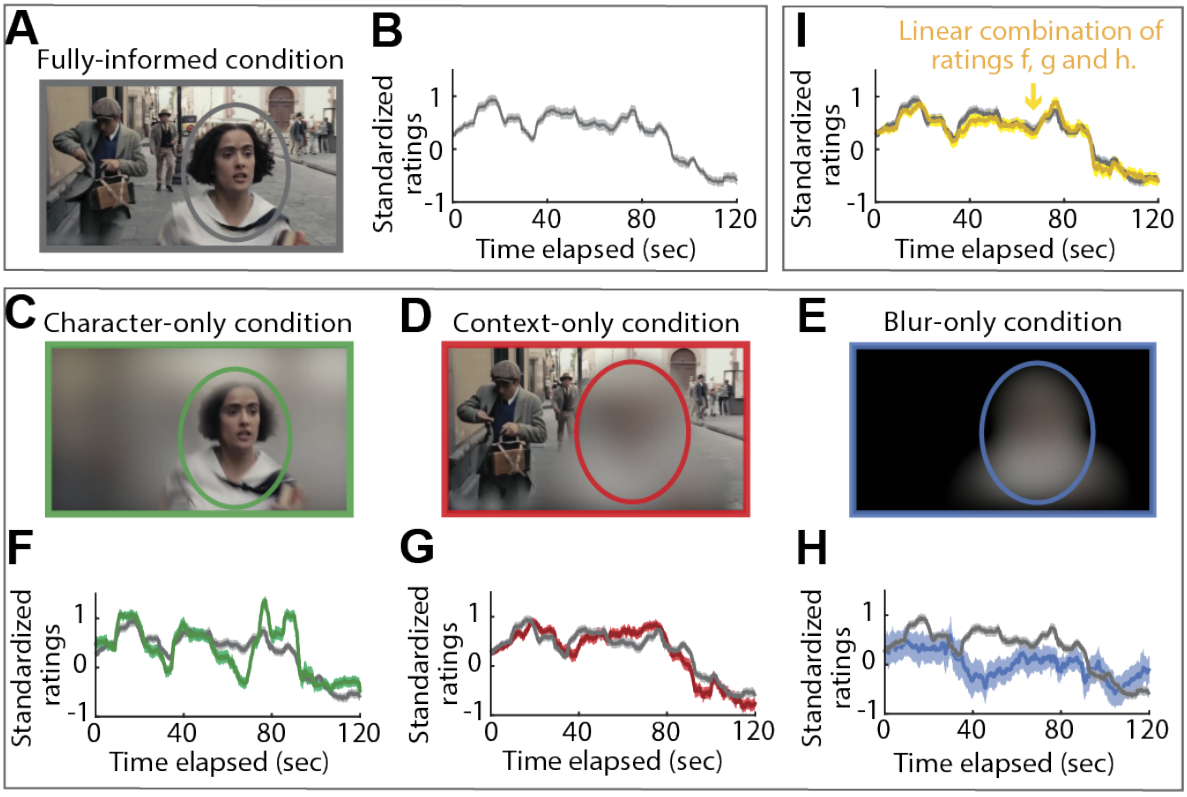


Fig. 2.3. Experiment 2. (A) Fully informed condition: tracking the affect of a visible target in a visible context (the female character in this particular example; circled in gray). (B) Example fully informed ratings of the target (gray curve). (C) Character-only condition: tracking the visible target (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.

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 (Fig. A3B in *Appendix A*). 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 (Fig. A3C in *Appendix A*).

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. 2.4A). 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. 2.4B, 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 (Fig. A4D in *Appendix A*) and across various basic emotion categories annotated by state-of-the-art computer vision models (Fig. A4E in *Appendix A*). 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. 2.4B, 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 Fig. A5 in *Appendix A*). In addition, the blur-only ratings contributed only 1.5% of the total variance explained (bootstrapped 95% CI: 0.69–2.5%; Fig. 2.4B, 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 non-significantly 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 *Linear Regression Analysis* in *Materials and Methods*, for a comparison with nonlinear models).

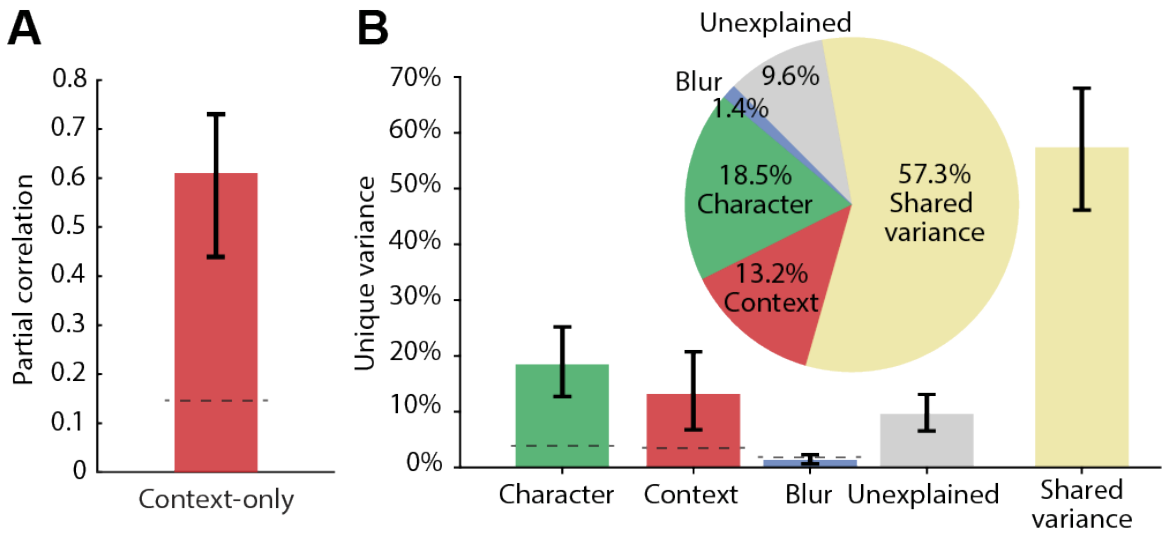


Fig. 2.4. (A) Mean partial correlations between context-only affect ratings and fully informed affect ratings of the target when controlling for the character-only affect ratings of the target. (B) Proportion of unique variance in the fully informed affect ratings that could only be explained by context-only 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 (Permutation Test in *Methods*).

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 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 Fig. A3 F and H in *Appendix A* 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 (Fig. A3 E and G in *Appendix A*).

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 Fig. A6A in *Appendix 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 Fig. A6B in *Appendix A*) 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 (Fig. A4D in *Appendix A*) and different basic emotion categories (Fig. A4E in *Appendix A*). Context is usually taken as having a modulatory influence (Barrett et al., 2011; Kayyal et al., 2015; Righart

& de Gelder, 2008a; Wieser & Brosch, 2012), although recent theories suggest that context might shape and influence the actual perception of emotion signals (Aviezer et al., 2017). 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 (Aviezer et al., 2011; Kret et al., 2013; Mavratzakis et al., 2016; Righart & de Gelder, 2008a; Righart & de Gelder, 2008b Whitney & Yamanashi Leib, 2016; Yamanashi Leib et al., 2016). 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 (Whitney & Yamanashi Leib, 2016). 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 (Gallese & Sinigaglia, 2011). 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 (Gopnik & Wellman, 1994). 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 (Fernández-Dols & Crivelli, 2013; Wenzler et al., 2016). 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 (Russell, 2003). 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 (Fig. A4E in *Appendix A*), 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 (Mayer et al., 2008) and emotion simulation (Zhou et al., 2017). Although the widely studied construct of emotional intelligence is highly debated (Mayer et al., 2008), 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 (Mayer et al., 2003; Russell & Dols, 1997) 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 (Dhall et al., 2017). 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 (Kosti et al., 2017). 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

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 (Brainard 1997; Pelli, 1997). 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. A2 in *Appendix A*).

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 (Kwiatkowski et al., 1992) 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.

Overall, Chapter 2 show that context provides a substantial and unique contribution to the perception of affect. However, it leaves open the question of how fast the contextual information becomes available or whether context is accessible as quickly as facial expressions. The goal of Chapter 3 is to test this.

Chapter 3: Inferential affective tracking reveals the remarkable speed of context-based emotion perception

Introduction

Rapid inference about the internal emotional states of others is an essential and unique human ability. It is necessary for understanding others, interpersonal communication, and adaptive social functioning. Impaired emotion understanding is associated with a number of disorders, ranging from autism to schizophrenia to depression (Kohler et al., 2004). The cognitive foundation of emotion understanding rests on our ability to derive and integrate information from a variety of cues across different modalities, including but not limited to facial expressions (e.g. Calder & Young, 2005; Ekman, 1992), body postures (e.g. de Gelder et al., 2015), background scenes (e.g. Chen & Whitney, 2019; Righart & de Gelder, 2008a), and vocal expressions (e.g. Cowen et al., 2019).

Previous research on emotion recognition has disproportionately focused on one channel of emotional information: the perception of facial expressions. This might be because faces are often thought to be attention-grabbing, uniquely salient, or evolutionarily significant (Ekman, 1992). The processing of facial expressions has also been considered to be rapid, efficient, and automatic (e.g. Fischer & Whitney, 2011; Poncet et al., 2019; Yang & Yeh, 2018). However, faces are usually encountered within situational contexts in everyday life, and humans can seamlessly integrate contextual information in the process of recognizing emotion. For example, recent studies show that human observers can readily learn and utilize associations between emotion context and neutral stimuli (e.g. Ventura-Bort et al., 2016). Sometimes, contextual information can even facilitate and speed up the processing of faces when information is limited or ambiguous (Falagiarda, & Collignon, 2019; Liedtke et al., 2017).

In recent years, an increasing body of research has shown that context can influence and modulate the interpreted emotions of facial expressions (Barrett et al., 2011; Wieser & Brosch, 2012) and this process has been suggested to be fast and automatic. For example, visual context can strongly influence perceived emotions from facial expressions when the context is irrelevant or subliminally presented (Aviezer et al., 2011; Mumenthaler & Sander, 2015). This context effect remains intact even when observers are cognitively loaded by a concurrent task (Aviezer et al., 2011). The magnitude of this context effect has also been found to correlate with the degree of enhancement of an early electrophysiological component (Meeren et al., 2005; Righart & de Gelder, 2006). Despite these advances, contextual information is still often regarded as secondary to facial information, mainly incorporated to modulate or disambiguate perceived emotion in faces.

Although there has been extensive research on perceived emotion from faces in the presence of contextual information, and the interaction between these sources of information, the inference of emotion from context alone in the absence of facial expression has remained largely unknown. However, a recent study shows that observers make remarkably good predictions of other peoples' affect (valence and arousal) when only the contextual information is available, while the face and body are blurred out (Chen & Whitney, 2019). Contextual information alone is therefore sufficient for an accurate interpretation of affective state, and the influence of the context on perceived emotion can be as substantial as the facial expression itself (Chen & Whitney, 2019). Context may therefore be a primary cue to emotion, not simply a secondary or modulatory cue.

The primacy and usefulness of context-based emotion perception depends on the speed of the available information. If context is as efficient (fast) as when the face and body information are present, it would suggest that context plays a pivotal and primary role. Previous literature has typically regarded the use of context alone, in the absent of overt expressions, to be indirect and deliberate because we have to rely on abstract causal principles rather than direct perceptual cues (Ekman, 1992; Skerry & Saxe, 2014). Further, conceptual models of perception tend to assimilate context or scene effects only at a relatively late and high-level processing stage (Bar, 2004). The implication of these previous studies is that contextual effects on emotion recognition might be relatively slow, or at least slower than the recognition of a facial expression. This may be, in part, because previous studies used unnatural or static stimuli and did not dynamically measure emotion.

In this paper, we adopted the inferential affective tracking method previously introduced in Chen & Whitney (2019) to characterize the continuous process of affect inference using dynamic and naturalistic stimuli. Here we developed a new analysis approach to measure the speed of recognizing emotion from contextual information alone. We further provide additional experiments to support and validate the precision of this approach. Our method is ideally suited to measure the speed of context-based emotion perception in the millisecond (msec) range, and it will be useful for quantitative models of emotion perception.

Experiment 1

The goal of experiment 1 was to quantify the speed of inferring emotion from contextual information alone, relative to the speed of recognizing emotion with all available information including facial expression. There are competing hypotheses in this experiment. One hypothesis suggests that contextual effects on emotion recognition might be relatively slow, in which case one might expect that inferring emotion from only contextual information would lag behind recognizing emotion with all available information. On the other hand, it is possible that contextual information could be processed with a short latency or in parallel with facial expression information, in which case the latency of inferred emotion perception could be very brief.

Method

Participants. We tested 90 healthy participants in total (16 male; age range 18 – 34, Mean = 21.0, SD = 4.22). They were students from the University of California, Berkeley participating for course credits. All participants had normal or corrected-to-normal vision. Informed consent was obtained from all participants and the study was approved by Institutional Review Board at the University of California, Berkeley.

Stimuli. We made use of a publicly available stimulus set from Chen & Whitney (2019). The stimulus set was originally made for experiments that assessed the unique contribution of context versus face and body information in emotion recognition. The stimuli consisted of 34 silent (no audio) video clips derived from Hollywood movies, home videos and documentaries. The lengths of the videos ranged from 36 seconds to 160 seconds, totaling 2,749 seconds for all videos. The resolution of the videos was 1280 x 720 and the frame rate was 30 frames per second. The original silent videos with all visual information visible were defined as the “fully informed” condition (Fig. 3.1A). For each video, we selectively masked the face and body of a target character, frame-by-frame, with a Gaussian blurred mask, to generate stimuli for the

"context-only" condition (Fig. 3.1B). In the context-only condition, the target character was never visible. The blurred mask appeared in the videos an average of 77.1% (SD = 19.1%) of the time. During the period when the blurred mask was present, the mask area covered an average of 31.5% (SD = 11.6%) of the entire frame.

To effectively evaluate tracking performance, we used video clips that contained variations in emotion. This means that many videos in our dataset contained more than 1 emotion and some transitions from positive to negative emotions or from negative to positive. In most of the clips, the affect was quite heterogeneous (Fig. B1 in *Appendix B*).

Procedure. Participants completed the experiment on a custom-made website online. The videos were presented to participants in a random order. Half of the videos that each participant viewed were from the fully informed condition and the other half from the context-only condition. Although it is slightly more complex, this mixed within-subject design has several advantages. First, every participant viewed a given video in either the fully informed or the context-only condition, but not both conditions, which avoided any memory or interference effects between conditions. Second, this design controlled for subject-specific sources of noise such as network latency or monitor settings. As a result of the random assignment of videos to different conditions for each subject, different videos in the same condition may have had a slightly different number of participants assigned (within +/- 1 participant). On average, for every video in either the fully informed condition or the context-only condition, we collected affect ratings from 45 participants. To record real-time affective judgments, a 2D valence-arousal affect rating grid was superimposed on top of the video (Fig. 3.1A & 3.1B). Participants were required to position the mouse at the center of the affect rating grid before the video presentation. As participants watched each video, and in real-time, they were instructed to move a mouse pointer within the affect rating grid to continuously report the valence and arousal of the (blurred or visible) target character in the video. The mouse position was recorded every 20 milliseconds (50 Hz). After a video ended, participants were asked whether they had seen the video prior to the experiment, and they rated their level of familiarity with the video clip on a scale from 1 (Not at all familiar) to 5 (Extremely familiar). Participants were also asked whether the video played smoothly.

To estimate the noise ceiling in our data, 52 participants were asked to rate a random subset of video clips twice. The second rating was collected approximately 1 hour after the first rating of the clip. The mean correlation coefficients between the initial ratings and the repeated ratings was used as the noise ceiling to normalize cross-correlation coefficients.

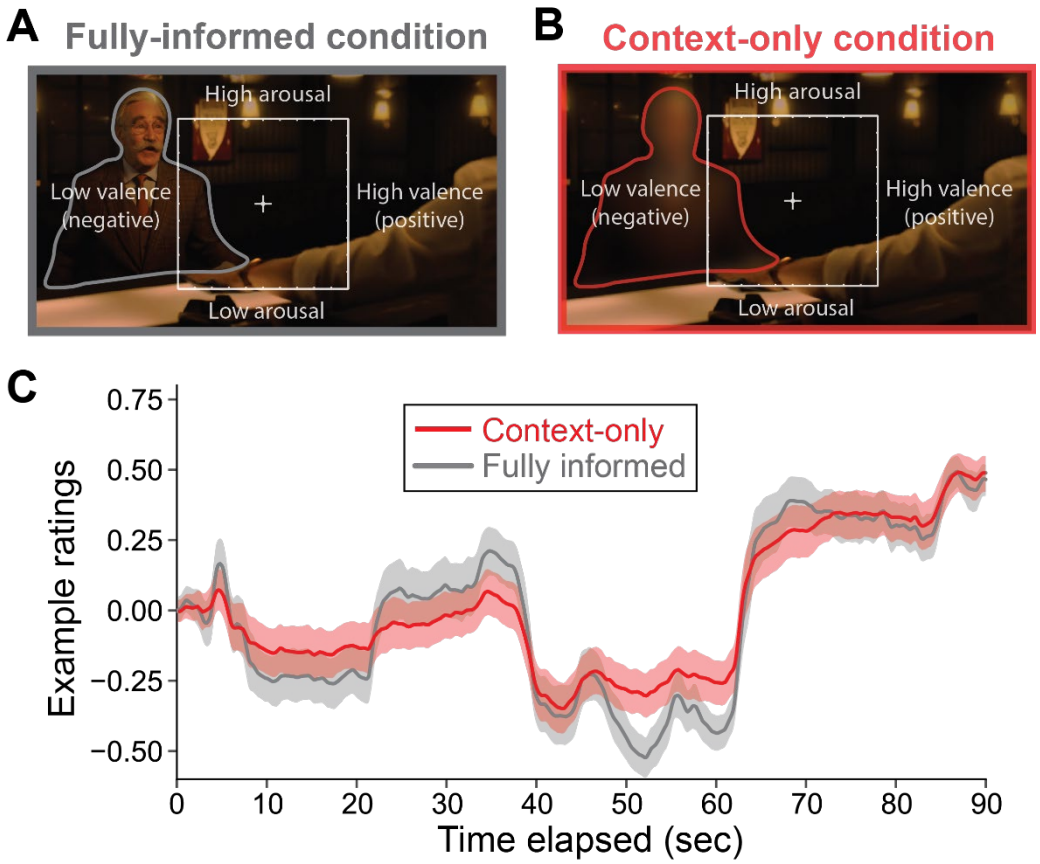


Fig. 3.1. Experimental conditions and data from a single example video in experiment 1. (A) Participants viewed a silent movie clip while continuously reporting the valence and arousal of a specified character in the video. In the fully informed condition, participants were asked to track the affect of the target character (outlined in gray) when everything was visible. Due to copyright restrictions, the example video frame here is for visualization purposes only; the full set of videos is available here: https://osf.io/f9rxn/. (B) In the context-only condition, participants tracked the blurred target (outlined in red) while the context remained visible. (C) Example raw context-only valence ratings of the invisible target (red curve, n = 41 participants) appear to follow a similar time-course as the fully informed valence ratings of the visible target (gray curve, n = 42 independent participants). The data here are for one single video; a total of 34 videos were tested. Shaded regions represent bootstrapped 95% confidence intervals.

Data analysis. We confirmed that participants reported smooth video playback in 98.4% of the trials and the remaining 1.6% of trials were removed from the analysis. We also confirmed the exclusion of these 1.6% of trials did not change the results. We obtained mean affect ratings for every video under each condition by averaging across responses from all participants. This helps reduce noise in the data that is caused by idiosyncrasies from individual participants. We then used a cross-correlation analysis to detect the time lag between the mean fully informed affect ratings and the mean context-only affect ratings. Many studies on perception, performance, psychophysiology, and neuroscience use time-lagged cross-correlation analysis to assess the similarity and synchrony relationship between pairs of time series or signals (Dean & Dunsmuir, 2016). Cross-correlation is measured by incrementally shifting one signal in time and repeatedly calculating the correlation between two signals (Fig. B2 in *Appendix B*). We used this

technique to compare the emotion inferred when all information is available with the emotion inferred when only context information is available. The temporal offset at which the two signals are most synchronized (correlated) indicates the difference in perceptual latency between the context-only and fully informed conditions.

Time series analysis requires the series to be stationary (Shumway & Stoffer, 2011), which is defined as having constant statistical (e.g mean and variance) properties that do not change over time. We transformed our time series data to be stationary by applying differencing (Sims, 1988), which involves subtracting every value x_t from x_{t+1} to obtain successive differences between adjacent values in time. The Dickey-Fuller test confirmed that all transformed/differenced time series were stationary.

For every video, we computed the cross-correlation function (CCF) between the context-only and the fully informed condition using the differenced affect ratings. We applied Fisher z transformation on the CCFs and averaged the transformed z values to obtain the mean Fisher z. transformed CCF. We then estimated the noise ceiling by computing the CCFs between initial and repeated differenced ratings made by the same subject. These CCFs were also transformed to Fisher z values and averaged to obtain the mean Fisher z transformed CCF across videos. The peak Fisher z value of the mean transformed CCF between initial and repeated ratings was identified as the noise ceiling. We then divided the mean Fisher z transformed CCF between the context-only and the fully informed condition by this noise ceiling, and inverse transformed the normalized Fisher z back to Pearson r values, in order to obtain the normalized mean CCF across all videos. We then fit a skew-Cauchy distribution (Bahrami et al., 2010) to the mean normalized CCF in order to capture the shape of the CCF. We confirmed that all skew-Cauchy curve fitting reached successful convergence to the optimal parameter values to ensure goodness of fit. We measured the time lag of the context-only condition by identifying the lag that has the highest correlation value along the skew-Cauchy curve fitted on the mean normalized CCF. The measured time lag was based on the mean CCF across videos because the focus of our study is on the temporal characteristics of context that are general to all video stimuli but are not specific to a single stimulus. The normalizing procedure by using the noise ceiling did not affect the result of time lag detection because the same noise ceiling was uniformly applied to cross correlation values of all time lags.

There is inevitable temporal dependency in the time series data because our method involved continuous affect ratings collected while viewing dynamic videos. Therefore, we did not use parametric tests (e.g. ANOVA and t-test) to evaluate statistical significance, because they make assumptions about the data and its distribution, which would often not hold true for our continuous data. Instead, we used non-parametric resampling (e.g. bootstrapping) and Monte Carlo permutation methods to generate null distributions and confidence intervals. To generate null distributions for CCFs, trial labels (whole continuous sequences of ratings) were shuffled (time points were not shuffled) and the CCF was calculated between affect ratings from randomly paired videos with different target characters. This permutation method preserved the exact temporal structure and autocorrelations inherent to the continuous ratings but not any clip-specific information. We estimated the bootstrapped confidence intervals by randomly sampling the CCFs of individual videos with replacement, recalculating the mean CCF across sampled videos and reidentifying the peak lag (that has the highest correlation value) from the bootstrapped mean CCF. This process was repeated 10,000 times to generate bootstrap distributions for the mean CCF and its peak lag; 95% confidence intervals were calculated based on the bootstrapped distributions.

As autocorrelation in time series could lead to spurious correlations, prewhitening has been proposed as a preprocessing step before calculating cross-correlation functions to further remove autocorrelation in time series (Shumway & Stoffer, 2011). Prewhitening is typically performed by fitting autoregressive integrated moving average (ARIMA) models to original time series and separating out the time series of residuals from the original series as the prewhitened series. Prewhitening has been shown to be effective when applied to some cases (Dean & Dunsmuir, 2016; Probst et al., 2012), while it is not always informative and can be detrimental in other cases (Bayazit & Onoz, 2006; Razavi & Vogel, 2017). In a separate analysis, we applied prewhitening to the differenced time series and confirmed that our basic results remained consistent whether or not it was applied.

Results

We calculated the skew-Cauchy fitted mean CCF between the context-only affect ratings of the invisible target character and the fully informed affect ratings of the visible target character (e.g., the cross correlation between the red and gray data in Fig. 3.1C). This reveals the relative delay between the use of contextual information alone and the use of all available information including the face and body. We observed a significant peak normalized cross-correlation between context-only and fully informed ratings (red line in Fig. 3.2A; mean r = 0.52, bootstrapped 95% CI: 0.38 – 0.65; $p < 0.001$, permutation tests), which confirmed that context information alone is indeed informative when inferring the affect of invisible characters. It is noteworthy that the fully-informed and context-only ratings were made by different groups of independent participants, so the cross correlation is not confounded by within-subject dependence or memory.

More importantly, we found that the peak of the skew-Cauchy fitted mean CCF had effectively zero time lag (Fig. 3.2A). We estimated the variability of the measured time lag using bootstrapping, which revealed no substantial lag (Fig. 3.2B and red shaded area in Fig. 3.2A, mode lag: 0 msec, mean lag: -33 msec, bootstrapped 95% CI: -240 to 60 msec). These results suggest that, on balance, context information alone is used nearly as fast as having additional facial expression information for emotion perception.

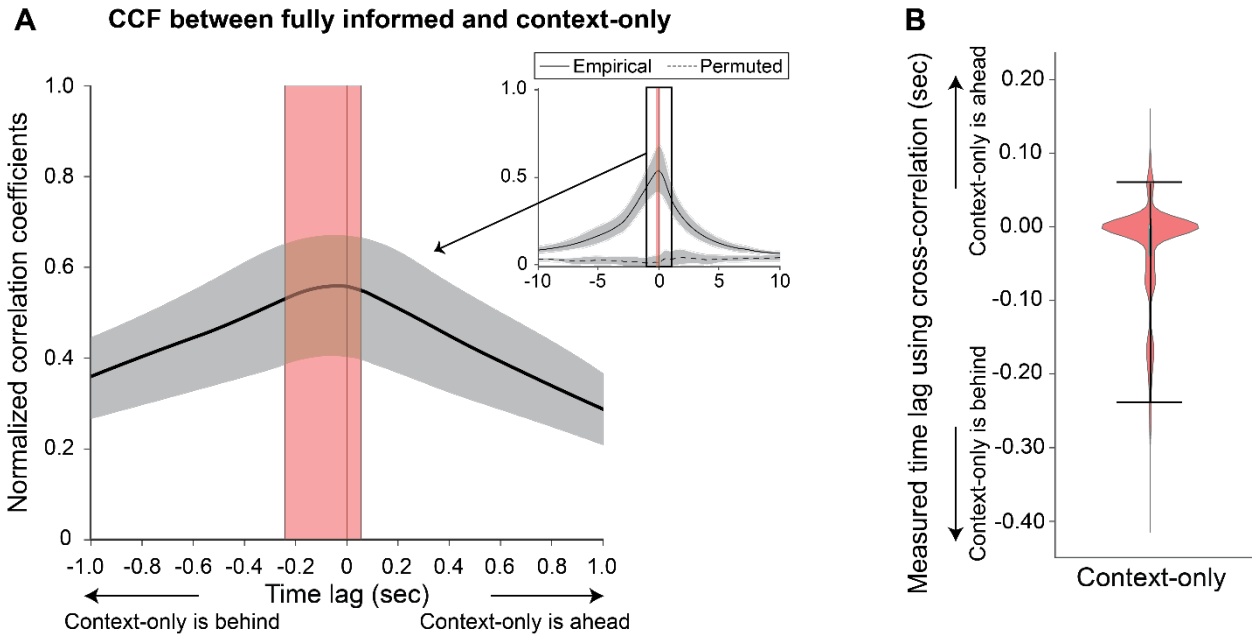


Fig. 3.2. Results of experiment 1. Skew-Cauchy fitted cross correlation functions (CCF) for detecting a time lag between conditions. (A) Skew-Cauchy fitted cross correlation functions (CCF), between context-only affect ratings of the invisible target and fully informed affect ratings of the visible target, as a function of the time displacement/lag between them (solid black line). The shaded gray region around the black lines represents bootstrapped 95% confidence intervals of mean cross correlation coefficients averaged across clips and targets. The red shaded area represents bootstrapped 95% confidence intervals of measured peak lags identified from the skew-Cauchy fitted CCF. The dashed line near (near zero, in the inset plot) represents the permuted null cross correlation functions generated by shuffling the video clip labels of continuous ratings. (B) A violin plot of the measured time lags in the context-only condition estimated by bootstrapping the peak of the CCF in panel A. The peak delay is narrowly tuned and clustered around zero lag. Error bars represent bootstrapped 95% confidence intervals of the mean measured time lag (same as the red region in panel A).

Experiment 2

The temporal cross-correlation analyses in experiment 1 revealed that using context information alone added no substantial processing delay compared to using information that includes the face. Contextual information is therefore sufficient to perceive emotion with very little if any delay. However, one might wonder whether our inferential emotion tracking method or the cross-correlation approach have enough precision to detect a time lag if it is indeed present. We designed experiment 2 to test the precision of our method for detecting the temporal lag in tracking emotion continuously. We inserted a small lag (100 msec in experiment 2a and 200 msec in experiment 2b) in the video stimuli to create physically lagged conditions. We expected to find a significant time lag when cross-correlating between the affect ratings of the lagged condition and those of the no-lag condition.

Method

Participants. In total, we tested 80 healthy participants in experiment 2a (18 male; age range 18 – 26, Mean = 20.2, SD = 1.48) and 76 healthy participants in experiment 2b (21 male;

age range 18 – 23, Mean = 20.1, SD = 2.43). Participants in experiment 1, 2a, and 2b did not overlap. They were students from the University of California, Berkeley participating for course credits. All participants had normal or corrected-to-normal vision. Informed consent was obtained from all participants and the study was approved by Institutional Review Board at the University of California, Berkeley. With this sample size and the statistical effect we observed in this study, we can reach a power of over 0.9 with an alpha value of 0.05.

Stimuli. The stimuli in the "no-lag" condition were identical to the fully informed condition in experiment 1 (Fig. 3.3A). In a separate condition, we edited these video stimuli by inserting a lag (100 msec in experiment 2a and 200 msec in experiment 2b), close to the start of each video; these were the "lagged" conditions (Fig. 3.3B). To insert the lag, we selected a random video frame between the first 5 to 10 seconds of each video, and we repeated the same frame for 100 (or 200) msec. The remaining frames in each video were therefore lagged by 100 (or 200) msec compared to the no-lag condition. Any fluctuations in timing could only add noise and reduce the measured precision but could not introduce a systematic lag between conditions.

Procedure. The procedure was identical to experiment 1. The videos were presented to participants in a random order, with half of them from the lagged condition and the other half from the no-lag condition. Experiment 2 also used a mixed within-subject design because every participant viewed the same video in either the lagged or the no-lag condition, but not both conditions. On average, for every video in either the lagged condition or the no-lag condition, we collected affect ratings from 40 participants in experiment 2a and 38 participants in experiment 2b. The mouse position was recorded every 20 msec (50 Hz) in experiment 2a and every 100 msec (10 Hz) in experiment 2b.

Data analysis. The affect ratings collected for the first few seconds before the 100 (or 200) msec lag was introduced were removed from the cross-correlation data analyses. The truncated affect ratings were processed in the same way as experiment 1. To examine the reliability of our lag detection method, we split the data in the no-lag condition into two halves and computed the cross-correlation between the mean ratings obtained from the two halves of data. We expected to find a narrowly tuned zero time lag in this (0 msec) condition. To further demonstrate the precision of our method, we quantified the relationship between measured time lag and physical time lag by fitting a linear regression function on every bootstrap iteration using the data from 0 msec, 100 msec and 200 msec. We expected the fitted linear regression function to have a slope close to 1 if the peak lag measured using our method matched the inserted physical lag.

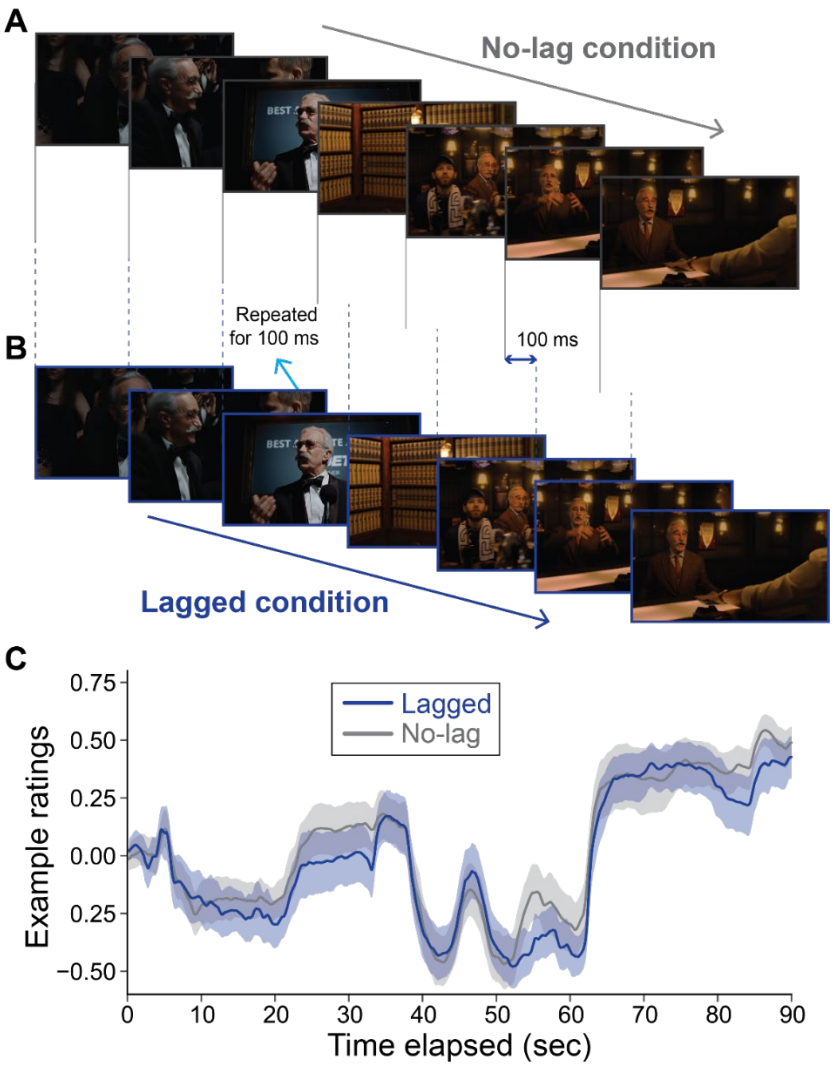


Fig. 3.3. Experiment 2a design and approach for an example video. (A) In the no-lag condition, participants viewed the original fully informed videos while tracking the valence and arousal of the target character. (B) In the lagged condition, participants viewed fully informed videos with a 100-msec time lag inserted at a random time after viewing the first 5-10 seconds of the video. To insert the time lag, we repeated the same video frame for 100 msec. As a result, all the video frames after the repeated frame lagged behind the no-lag condition by 100 msec. Experiment 2b had the same experimental design as experiment 2a except that the lag inserted was 200 msec. (C) Example raw lagged valence ratings of the target (blue curve) relative to the no-lag valence ratings of the target (gray curve), for one example video. The example ratings for the no-lag condition have data from 35 participants and the example ratings for the lagged condition have data from 36 participants. Shaded regions represent bootstrapped 95% confidence intervals.

Results

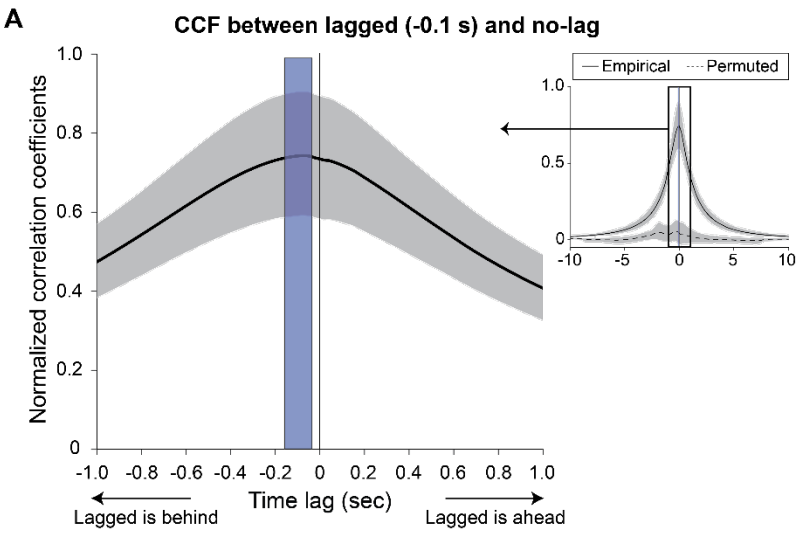
We calculated the skew-Cauchy fitted mean CCF between the no-lag affect ratings of the visible target character and the lagged affect ratings of the same character. We confirmed that the peak normalized cross-correlations between no-lag and lagged ratings were high for both experiment 2a (peak of the black line in Fig. 3.4A; mean: 0.75; bootstrapped 95% CI: 0.58 –

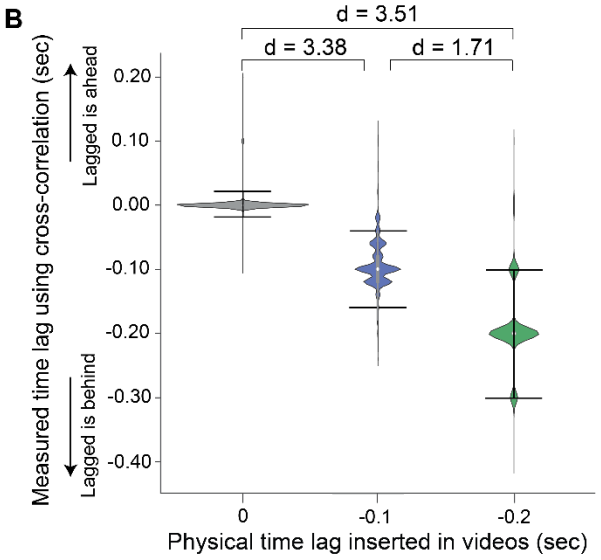
0.86; $p < 0.001$, permutation tests) and 2b (mean: 0.76; bootstrapped 95% CI: 0.56 – 0.87; $p < 0.001$, permutation tests).

More importantly, the peak of the skew-Cauchy fitted mean CCF is clearly shifted from zero time lag (see the peak of the black line in Fig. 3.4A). For experiment 2a, where we inserted a 100 msec time lag in the videos, we detected a significant time lag in the affect ratings (mode: -100 msec; mean: -87 msec; bootstrapped 95% CI: -160 to -20 msec; minus sign represents a lag instead of a lead; see the blue violin plot in Fig. 3.4B). In experiment 2b, where we inserted a 200 msec time lag in the videos, we detected a significant time lag in the affect ratings (mode: -200 msec; mean: -193 msec; bootstrapped 95% CI: -300 to -100 msec; the green violin plot in Fig. 3.4B). When we split the no-lag ratings into two halves (Monte Carlo) and compared them using cross-correlation (the 0 msec condition), we verified that the measured time lag was narrowly tuned to zero (mode: 0 msec; mean: -4 msec; bootstrapped 95% CI: -20 to 20 msec).

To show that our method can distinguish a 100 msec lag from a 200 msec or a 0 msec one, we quantified the effect size (Cohen's d) for the difference in measured time lags between lag conditions (Fig. 3.4B). We found a Cohen's d of 3.38 between the 0 msec and the 100 msec lag conditions, a Cohen's d of 1.71 between the 100 msec and the 200 msec lag condition, and a Cohen's d of 3.51 between the 0 msec and 200 msec lag condition. These Cohen's d values all indicate very large effect sizes (Cohen, 2013). Furthermore, we found that the linear regression function fitted on the bootstrapped distributions of 0 msec, 100 msec and 200 msec data has a mode slope of 1 (the black diagonal fitted regression line in Fig. 3.4C), which shows that that the measured time lag using our method matches the physical lag inserted. These results suggest that our method can resolve lags as small as 100 msec or less with high precision.

We also compared the measured time lag in the context-only condition in experiment 1 (Fig. 3.2B) to that of the 0 msec, 100 msec and 200 msec lag conditions in experiment 2 (Fig. 3.4B). As a reminder, the measured time lag in the context only condition was near zero (Fig. 3.2B). We found a Cohen's d of 0.75 between this context-only condition and the 100 msec lag condition, and a Cohen's d of 1.77 between the context-only condition and the 200 msec lag condition. These Cohen's d values indicate medium to large effect sizes (Cohen, 2013). Because we have obtained the bootstrapped distribution of measured lags for 0 msec (Fig. 3.4C, gray dot cloud), 100 msec (Fig. 3.4B, blue dot clouds) and 200 msec (Fig. 3.4C, green dot clouds) lag conditions, we computed the Bayes factor to evaluate which of these distributions fits the context-only distribution (Fig. 3.4C, red dot cloud) the best. We found a Bayes factor (BF_{12}) of 5.11 in the comparison between the 0 msec (H_1) and the 100 msec (H_2) distributions, and a Bayes factor (BF_{12}) of 9.78 in the comparison between 0 msec (H_1) and the 200 msec (H_2) distributions. These Bayes factors suggest moderate (Lee & Wagenmakers, 2013) to substantial (Jeffreys, 1961; Kass & Raftery, 1995) statistical evidence in favor of the 0 msec (H1) distribution compared to the 100 msec or the 200 msec (H2) distributions. Although the labelling of Bayes factors varies slightly across different references (Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2013) , we opt to focus the literal interpretation of the Bayes factors, the fact that the measured time lag in context-only condition is 5.11 times more likely to be under to the 0 msec lag distribution than the 100 msec lag distribution. The Bayes factors show that latency of context-only affect perception is much more likely to be drawn from a population with 0 msec lag than a population with 100 or 200 msec lag, confirming that the context information is available with a remarkably short latency.


A
CCF between lagged (-0.1 s) and no-lag
Normalized correlation coefficients
1.0
0.8
0.6
0.4
0.2
0
-1.0 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1.0
Time lag (sec)
Lagged is behind
Lagged is ahead
Empirical
Permuted
1.0
0.5
0
-10 -5 0 5 10


B
d = 3.51
d = 3.38
d = 1.71
Measured time lag using cross-correlation (sec)
Lagged is ahead
Lagged is behind
0.20
0.10
0.00
-0.10
-0.20
-0.30
-0.40
0
-0.1
-0.2
Physical time lag inserted in videos (sec)

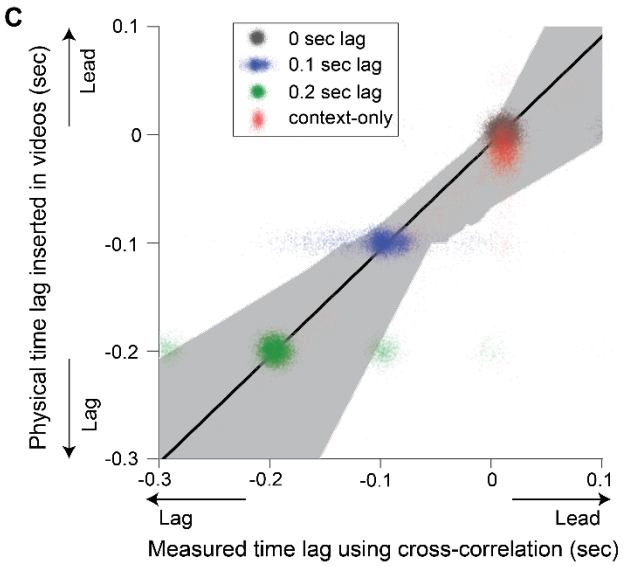

C
0 sec lag
0.1 sec lag
0.2 sec lag
context-only
Physical time lag inserted in videos (sec)
Lead
Lag
0.1
0
-0.1
-0.2
-0.3
-0.3 -0.2 -0.1 0 0.1
Lag
Lead
Measured time lag using cross-correlation (sec)

Fig. 3.4. Results of experiment 2. (A) Experiment 2A: skew-Cauchy fitted CCF (solid black line) between lagged (-100 msec) affect ratings of the target and no-lag affect ratings of the target as a function of the time displacement/lag between them. Blue shaded area represents bootstrapped 95% confidence interval of measured time lags identified from the skew-Cauchy fitted CCF. The dashed line (inset) represents the permuted null cross correlation functions generated by shuffling the video clip labels of continuous ratings. Shaded gray region around black lines represents bootstrapped 95% confidence intervals of mean cross correlation coefficients averaged across clips and targets. (B) Violin plots of the measured time lags for experiment 2a (-0.1 sec physical lag in blue), experiment 2b (-0.2 sec physical lag in green), and the split-half analysis (0 sec physical lag in gray). Error bars represent bootstrapped 95% confidence intervals of the mean measured time lag (same as the blue region in panel A). (C) The fitted linear relationship between physical time lag inserted in videos and measured time lag using our cross-correlation method. Gray shaded region shows the 95% confidence intervals of fitted linear regression functions using the bootstrapped data from 0 msec, 100 msec lag and 200 msec lag conditions. We then predicted the physical time lag of the context-only condition using the fitted linear regression function and the measured time lag of the context-only condition from experiment 1 (in red). The measured and the predicted physical time lag for the context-only condition are both located closer to the measured time lags of the 0 msec condition than the 100 or 200 msec lag conditions. Data points are jittered for visualization purposes.

Discussion

Our results provide the first measure of the speed of context-based dynamic emotion perception and show that the context is processed with a remarkably short latency, essentially as fast as using all available information including facial expressions. Our continuous inferential affective tracking technique (IAT), in combination with cross-correlation analysis (experiment 1), has high precision in detecting a sub-second temporal lags as established by empirical experimental manipulations (experiment 2). These results contrast with previous theories of inferred emotion from context, which implicitly or explicitly suggest that perceiving emotion from context is slower than emotion directly from faces (Skerry & Saxe, 2014; Bar, 2004). Our results support the alternative view that emotion from dynamic contextual information might be automatic and immediate. Seemingly complex context-dependent emotional inference and recognition is far more efficient than previously assumed.

The dynamic tracking method itself does not set limits on the precision of detecting a time lag between conditions. Although participants' affect ratings could be sluggish because of a large latency in the motor movement or mouse kinematics, this applies to all conditions (fully informed, and context-only). These sources of latency and temporal blurring do not affect our speed measurement because we focused on the difference between conditions, and the motor latency, for example, would therefore cancel-out in the comparison. Despite the 10 and 50 Hz sampling rate of the behavioral data, comparing between conditions can reveal a reliable temporal lag of 100 msec or less. Averaging across trials, for example, allows one to measure temporal differences in reaction time much finer than the resolution of the device itself (e.g., Donders, 1969).

Although the current study only concerns valence and arousal, our tracking method has also been demonstrated with ratings in discrete emotion categories rather than affect (Chen & Whitney, in press). The tracking method was adapted from IAT and it was called inferential emotion tracking (IET). With IET, we showed that context remains essential for emotion

recognition regardless of whether the emotion is reported as dimensional or categorical. Based on the current study, it is therefore plausible that context information is used very quickly for both affective and categorical emotion perception.

It is worth noting that the striking findings in this study are related to the failure to demonstrate a consistent lag between conditions different from zero. It is unquestionable that one cannot conclude that the null hypothesis is true when one fails to reject it. This is where experiment 2a and 2b come in, to guide our interpretation of the measured CCF lag in experiment 1, by showing that our method can reliably detect a sub-second lag. One possibility is that the time lag between fully informed and context-only affect ratings is smaller than the limit that we have tested (< 100 msec). Another possibility is that experiment 1 involves a comparison between different conditions, which is inherently noisier than a comparison within the same condition. With the above constraints considered, we can still safely conclude that emotion inferred from context information alone does not yield a time lag as strong and consistent as the 100 msec lag tested in experiment 2a. Context information is therefore available as early as any emotion-related visual information, and it has a very fast influence on perceived affect.

Although our study has shown that inferring emotion from only contextual information is processed with little delay, our results do not speak to what specific information in the context-only condition is essential for such fast emotion inference. The information in blurred masks on its own contains some color or residual outline motion, but that cannot be used to perceive emotion accurately, as previously demonstrated (Chen & Whitney, 2019). However, the blurred mask is embedded in the scene context and it may interact with the context in a way that provides useful information. Scenes with other characters as part of the context may provide more information to allow for faster inference of emotion. However, we did not find a large or easily interpretable difference in measured time lag between videos with one character only or with more than one character (Fig. B3 in *Appendix B*). The overall valence of the affect ratings also did not change the results much: the relative latency of inferential emotion tracking was near zero for videos that were relatively more positive or negative (Fig. B4 in *Appendix B*). One might expect that familiarity with video content might play a role in determining the lag. We therefore analyzed the data excluding trials in which participants reported familiarity with the video. We found that the IET latency was similar regardless of familiarity with the videos (Fig. B5 in *Appendix B*).

Our findings are consistent with and extend a large body of work showing that the perceptual organization and integration of visual contextual information is fast and automatic. There are many extensively investigated visual processes that require the integration of some form of visual context and they have been suggested to be pre-attentive and automatic (for a review see Albright & Stoner, 2002). These rapid processes include but are not limited to analysis of shadows (Rensink & Cavanagh, 2004), perceptual filling-in (Mattingley et al., 1997), texture segmentation (Zhaoping, 2000), figure-ground segregation (Kimchi & Peterson, 2008), etc. The perception of facial attributes such as expressions and attractiveness is also influenced by the context of other faces presented in the recent past or simultaneously (e.g. Liberman et al., 2018; Wedell et al., 1987). Some studies have shown that neurons at early stages of cortical processing are involved in detecting contextual cues and representing the modulated information (Albright & Stoner, 2002). This evidence suggests that context-based processing could be primitive and efficient. Our results extend this to the some of the highest levels of visual cognition, including inferential emotion perception.

Our findings speak to the question of what enables emotion inference from context to be so fast. Context could be faciliatory for emotion recognition in a temporal and spatial manner. Emerging evidence in neuroscience supports that the human brain performs mental inference based on predictive encoding (Friston, 2010). This account has subsequently been extended to social and emotion perception (Otten et al., 2017). According to this account, extracting sequential regularities embedded in the temporal context to form predictions about upcoming events is an essential cognitive and neural process. These sequential regularities in the recent past can serve as the temporal context to constrain and shape inferences of others' emotions (Kimura et al., 2012). Similarly, spatial context may contain heuristics based on behavioral regularities in the social environment, which can then provide shortcuts for emotion processing (Marsh, 2002). For example, emotions like panic tend to spread in crowds, and get intensified beyond what any individual face can signal. Furthermore, detailed contextual information may facilitate the understanding of others' emotional states by actively engaging other empathic and interoceptive processes (Melloni et al., 2013).

Our results have implications for the underlying neural mechanisms of emotion perception. A common view is that contextual information modulates the neural processing of facial expressions through feedback connections (Wieser & Brosch, 2012). Our study suggests an alternative possibility, albeit speculative, that there might be a parallel pathway, independent of the pathway of facial analysis, for processing and extracting affective information from visual background context. This context-pathway is supported by work showing that independent and unique variance in emotion perception is carried by face and contextual information (Chen & Whitney, 2019) and it would likely involve brain regions that support the analysis of objects, scenes, bodies, and actions that constitute the interpretation of visual context. Using our approach, future neuroimaging experiments could generate encoding and decoding models to isolate the neural mechanism of context-based emotion perception.

Taken together, our findings reinforce the idea that context plays a critical role in supporting the rapid and robust understanding of others' emotion. This has practical implications for affective computing, which stresses the importance of fast and accurate emotion recognition. In light of our results, the implementation of a context processing stream should not be regarded as peripheral and superfluous, but rather essential. Emotion inference from context is a seemingly complex and challenging problem as visual scene context is heterogenous and the processing of it seems computationally expensive. However, we have shown that the human brain resolves it with remarkably speed and efficiency in a relatively effortless manner. To understand and exploit the brain's full potential in the realm affective computing, it is therefore important to shift our focus towards studying the cognitive and neural mechanisms underlying context-based emotion perception.

Overall, Chapter 3 show that contextual information is processed with a remarkably short latency, essentially as fast as using all available information including facial expressions. However, the IAT method used in Chapter 2 and 3 only characterizes the influence of context on the affective dimensions of valence and arousal but not categorical emotions. The goal of Chapter 4 is to address this by extending the IAT method to examine the role of context in categorical emotion perception.

Chapter 4: Inferential emotion tracking (IET) reveals the critical role of context in emotion recognition

Introduction

Facial emotion expressions are widely studied and provide important cues about one's emotional state. But to the extent that faces are not seen isolated in real-life, and they are often accompanied by a variety of other contextual cues, we rely on more than just facial features to perceive emotion. Context is important for emotion recognition and there has been a growing body of literature showing that contextual information apart from facial expressions (e.g. words, body postures, visual scenes) modulates the perception of emotion (Aviezer et al., 2017; Barrett et al., 2011; Betz et al., 2019; Chen & Whitney, 2019; de Gelder & Van den Stock, 2011; Wieser & Brosch, 2012). Visual scene context contains abundant emotion-relevant information that human perceptual systems are sensitive to and can readily use to infer emotion. However, previous research on the role of visual scene context typically used unnatural pairings of facial expressions with independent background information, and they only investigated a very limited range of scenarios and emotion categories (Aviezer et al., 2012; Barrett & Kensinger, 2010; Kayyal et al., 2015; Kret & de Gelder, 2010; Reschke et al., 2019). A recent study addressed these limitations and demonstrated the enormous importance of visual context using naturalistic and dynamic videos across a wide range of situations (Chen & Whitney, 2019). This study introduced the "inferential affective tracking" (IAT) method and showed that the context is sufficient for recognizing the affect (valence and arousal) of a character in the video even when the face and body of the character is made unavailable. Beyond information available from facial expressions and body postures, visual scene context also provides necessary and unique information for accurately perceiving affect over time.

A limit of the IAT method is that it only characterizes the influence of context on the affective dimensions of valence and arousal. Yet, this dimensional representation of affect is not the same as discrete emotion categories. Indeed, the dimensional and categorical approaches for characterizing emotion have been considered sufficiently different that they have been contrasted against each other (Cowen & Keltner, 2017; Kragel & LaBar, 2016; Russell, 2003). On a theoretical basis, emotion dimensions like valence and arousal have been thought as the core to all affective experiences; they are relatively primitive, raw, and disconnected from specific emotion categories (Clore & Ortony, 2013; Russell, 2003). In contrast, emotion categories might represent more detailed or subtle variations in emotion that might not be captured by affective dimensions (Cowen & Keltner, 2017). For example, distinct emotion categories like anger and fear might have similar values in terms of valence and arousal (Bradley & Lang, 1999; Warriner et al., 2013). Regardless of whether emotions are truly categorical in nature or not, it is important to understand emotion in a categorical context because people regularly express and recognize emotional states in terms of discrete categories in daily social interaction. Although the inferential affective tracking technique showed that affect recognition requires context (Chen & Whitney, 2019), it remains to be tested whether context is necessary to perceive discrete emotion categories in natural dynamic scenes.

Although there is a wealth of research on basic emotion categories and even some work on the modulating effect of context on emotion category perception (Aviezer et al., 2012; Calbi et al., 2017; Meeren et al., 2005; Kret & de Gelder, 2010; Reschke et al., 2019), it is not commonly assumed in emotion research that context would or should play a critical role in the perception of discrete emotion categories in dynamic natural scenes. Discrete emotion categories

are often considered to be expressed with certain facial features and movements (Cordaro et al., 2018; Ekman, 1992; Matsumoto et al., 2008). Many studies and review articles give the impression of a one-to-one mapping between face and emotion by explicitly linking a single, unique facial configuration to each emotion category (for a discussion of this, see Barrett et al., 2019). And, in many cases, the operational definition of an emotion category is tantamount to a particular facial expression. Although some researchers have challenged the idea of unique face-emotion mappings and acknowledged that every emotion category can be expressed with a number of different facial configurations (Barrett et al., 2019; Keltner & Cordaro, 2015), it remains to be learned the extent to which context drives the inference of emotion categories from facial expressions.

The current study introduces a new paradigm that extends previous studies by demonstrating the unique and critical contribution of visual context in the categorical perception of emotion. Specifically, we studied how observers recognize the emotion of target characters in video clips collected from various sources including Hollywood movies, home videos, and documentaries. Within this context, we operationalized emotion as the affective mental states conveyed by the target characters that could be reliably categorized into discrete emotion categories. To quantify the contribution of context, we adapted the inferential affective tracking (IAT) method from Chen and Whitney (2019) and developed it to test categorical emotion perception rather than affect. This method is very similar to the IAT technique and so we call it "inferential emotion tracking" (IET). Using IET, our current study demonstrates that context is often necessary to most accurately perceive emotion categories, even when face and body information is available. Simply put, both context and character (face and body) information are essential to accurately identify emotion categories.

Method

Participants. In total, we tested 204 healthy participants (57 male, age range 18 - 45, Mean = 21, SD = 3.4). Our participants comprised of university students at the University of California, Berkeley participating for course credits. All participants were naive to the purpose of the experiment. The study was approved by Institutional Review Board at the University of California, Berkeley and informed consent was obtained from all participants. Participants were assigned to different experimental conditions randomly, resulting in an average of 42 independent participants (30 female and 12 male) in every condition. With this sample size and the statistical effect we observed in this study, we can reach a power of over 0.9 with an alpha value of 0.01.

Stimuli. The same set of video stimuli were used in a previous study to collect ratings of valence and arousal (Chen & Whitney, 2019). The videos were gathered from online video-sharing website 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. We chose the videos to portray a wide range of social situations (e.g. roadway interactions, interview, courtroom, farewell, competition, wedding, gift unwrapping, birthday party). We also balanced the number of videos with positive and negative emotions, as well as emotions of both high and low arousal.

The stimuli consisted of 33 silent video clips collected from various sources including Hollywood movies, home videos and documentaries. These videos portrayed a diverse range of situations: 12 Hollywood movie clips focused on interactions between multiple people, 9

Hollywood movie clips showed a single character alone with no interpersonal interaction, and 12 non-Hollywood videos from home videos or documentaries. Eleven out of the 12 home videos and documentaries included interpersonal contexts. The lengths of the videos ranged from 36 seconds to 160 seconds.

To record real-time emotion judgements, we designed an emotion rating circle and superimposed it on top of the video (Fig. 4.1A). Different locations within the circle represented different emotion categories. As they watched each video, and in real-time, participants were instructed to move the mouse to point to emotion categories that the target character appeared to experience. We chose 10 emotion categories, along with a neutral category, to display in the emotion ratings circle. We included the commonly studied six basic emotions (happy, sad, angry, fearful, disgusted, surprised; Ekman, 1992). As the 6 basic emotions is heavily skewed towards negative emotions with relatively high normative arousal, we added 4 other commonly studied emotion categories in order to achieve a relatively even distribution over normative valence and arousal. The non-basic emotion categories include boredom (Rozin & Cohen, 2003), calmness (Cowen & Keltner, 2017), contentment (Keltner & Lerner, 2010), and excitement (Shiota et al., 2017; Fredrickson, 1998). To make the task more intuitive, we arranged the emotion categories to satisfy the following criteria based on the affective norms of these emotion categories (Bradley & Lang, 1999): 1) 'neutral 'was always at the middle of the circle; 2) all emotions with positive valence were on one side (left or right for a given participant) and all emotions with negative valence were on the other side. For every participant, we randomly assigned either the left or right half to display the positive side of the emotion rating circle; 3) the normative arousal of the emotion categories decreased in order from top to bottom or from bottom to top, which was randomly assigned for a given participant.

For each video clip, we used a Gaussian blurred mask to selectively occlude specific visual information frame by frame to create different experimental conditions (Chen & Whitney, 2019). The original videos with everything visible was nominally defined as the fully informed condition (Fig. 4.1B). To create stimuli in the context-only condition (Fig. 4.1C), we used state-of-the-art object segmentation algorithms (Mask R-CNN; He et al., 2018) and video editing software (Adobe Premiere Pro CC) to selectively mask out a chosen target character in the video so that every detail of the character's face and body became invisible. These masks were then inverted to mask out all contextual background information leaving only the target character visible to create video stimuli for the character-only condition (Fig. 4.1D). To control for the residual information (e.g. shape, color or biological motion) accessible from the blurred target character, we kept the blurred target from the context-only condition and replaced the other regions with black pixels to create the blur-only condition (Fig. 4.1E). Our library of videos and emotion ratings have been made available at https://osf.io/46rtw/.

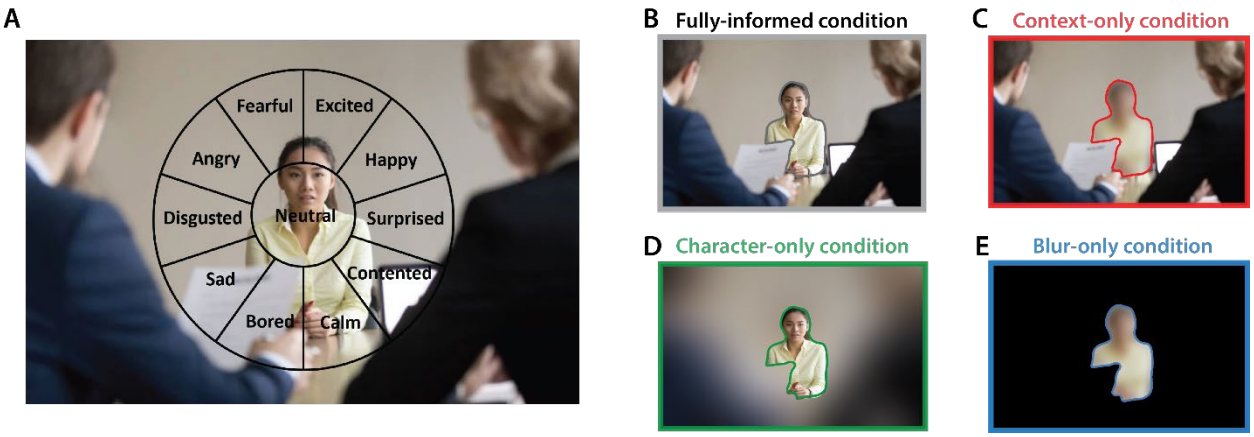


Fig. 4.1. Experimental paradigm. (A) Participants viewed a silent movie clip while moving a mouse pointer within the emotion rating circle (superimposed on the video) to continuously report the emotion of a chosen character in the video. (B) In the fully informed condition, participants were asked to track the emotion of the target character (the female, outlined in gray) when everything was visible. (C) In the context-only condition, participants tracked the blurred target (outlined in red) while the context remained visible. (D) In the character-only condition, participants tracked the visible target (outlined in green) while the context was blurred. (E) In the blur-only condition, participants tracked the blurred target (outlined in blue) while the context was masked completely by black pixels. Photo: reuse license purchased from iStock by Getty Images.

Procedure. We used a similar procedure as Chen and Whiney (2019). Participants completed the experiments on a custom-made website online. In one experiment, 126 participants were randomly assigned to view videos in one of the three conditions: context-only, character-only, and fully informed condition. In a second experiment, 79 participants were assigned to view videos randomly sampled from two conditions in order to keep participants engaged: 2/3 of the videos were from the blur-only condition and 1/3 of the videos in the context-only condition. In both experiments, observers were instructed to track and rate, in real time, the emotion of the target character (blurred or visible) while the video was playing. All video clips were presented in a random order. To familiarize with the task, participants completed a 2-min practice trial before starting the main experiment. Prior to starting a video, we informed participants of the identity of the target character by showing a frame containing the target character's face and body. In the context-only and blur-only conditions, this target character picture was blurred to avoid revealing any affective information. The mouse pointer was always centered on the 'neutral 'category when a video started, and mouse position was recorded every 100 ms (10 Hz) while the video was playing. In all data analyses, we excluded ratings collected within 3 seconds from when the video started playing. After a video ended, participants were asked whether they had seen the video prior to the experiment, and they rated their level of familiarity with the video clip on a scale from 1 (Not at all familiar) to 5 (Extremely familiar). We assessed whether participants lapsed or were non-responsive by calculating the longest duration that the participant had kept the mouse pointer in any single location. If the duration was longer than 20 seconds, the participant was reminded to pay more attention in future trials. In all other trials, the participant was given positive feedback.

Results

We aggregated ratings across all participants for each condition and calculated the percentage of participants who chose every one of the emotion categories at every sampled time point (see example data in Fig. 4.2A). To quantify consensus or between-subject agreement, we calculated the proportion of participants who chose the most selected emotion category for each time point (excluding the 'neutral 'category, the default mouse position).The relative frequency of the most selected emotion category in all video stimuli was highly correlated with the relative frequency of the corresponding affective English word reported in previous studies ($r = 0.907$, $p < 0.001$; Bradley & Lang, 1999), which supports the representativeness of our video stimuli (Fig. 4.2B; see Fig. C1 in *Appendix C* for a comparison between experimental conditions). Our video stimuli often displayed dynamic content with emotions changing over the time course of a video. On average, every video displayed about 4 different discrete emotions, and the dominant emotion was present for about 55.5% (SD: 20.0%) of the video duration.

In the absence of face and body information, participants agreed with each other about the emotion of invisible target characters when given available visual context information. If participants responded randomly, we would expect on average 10% (1 out of 10 non-neutral emotion categories) of participants to agree on the most selected emotion at any one moment. The blur-only condition is consistent with this, showing a distribution that peaks and is centered at the chance level agreement (Mean: 10.6%; SD = 4.18%; Fig. 4.2B, blue distribution). The between-subject agreement in other conditions was substantially different from the blur-only condition, indicating that participants agreed more frequently than chance (K-S test, $p < 0.001$, mean K-S statistic = 0.920). In the context-only condition when face and body information were masked and invisible, 98.2% of the agreement distribution falls above the chance level (Fig. 4.2B, red distribution), which is comparable to the fully informed condition (98.1%; Fig. 4.2B, gray distribution) and the character-only condition (97.4%; Fig. 4.2B, green distribution). These results suggest that when participants were only given contextual but not face and body information of the target character, emotion recognition remained robust without compromising between-subject agreement. Visual context can be sufficient for inferring emotions that can be shared among perceivers.

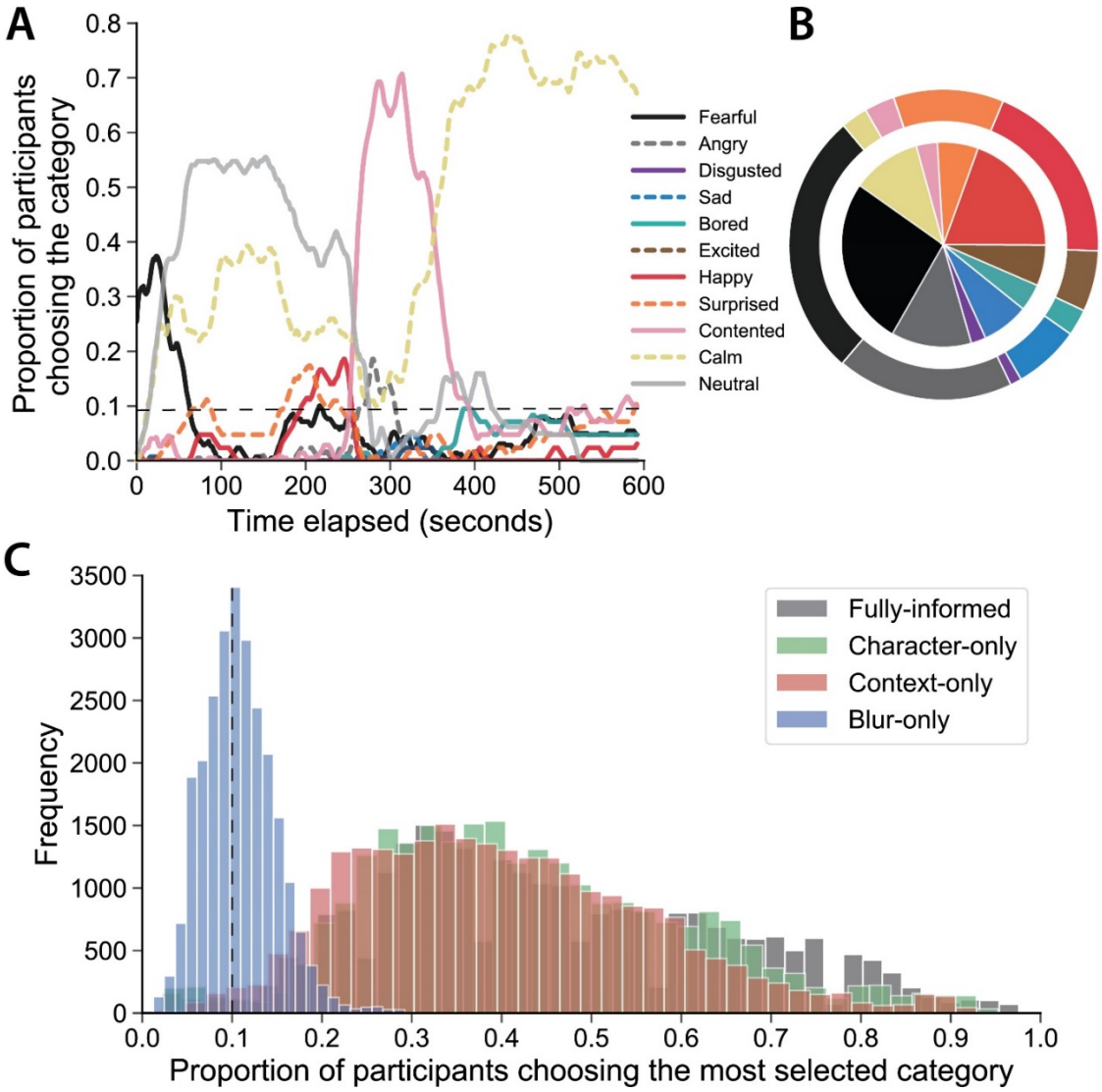


Fig. 4.2. (A) Categorical emotion ratings for an example video stimulus. The dashed horizontal line is chance level (1/11 ~ 0.091). (B) Inner pie: relative frequency of the most selected emotion category at all time points across all video stimuli. Outer pie: relative frequency of affective English words corresponding to the 10 categories in our emotion rating circle (Bradley & Lang, 1999). The color scheme is the same as panel A. The correlation between the reported emotion categories in our movie stimuli and the corresponding English word frequency was 0.907 ($p < 0.001$) (C) Between-subject agreement in categorical emotion ratings across all videos. The percentage of people who chose the most selected emotion category in each 100 millisecond of the videos (not including 'neutral'; the default mouse position) is shown. Chance level (1/10 = 10%) is indicated by the dashed vertical line.

Is the context necessary to perceive and track emotion accurately, even when face and body information are already available? To answer this question, we compared the most selected emotion category for all time points across different conditions. We aggregated ratings across all participants and preserved only the emotion category that was selected by the most participants for every sampled time point. To assess the unique contribution of context, we computed the

proportion of time in each video when only the most selected emotion in the context-only condition but not any other condition matched that in the fully informed condition. Similarly, we computed the proportion of time in each video when only the character-only condition or the blur-only condition exclusively matched the fully informed condition. The proportion of time when none of the conditions had a reported emotion that matched the fully informed condition is called 'unexplained' and the proportion time when the most selected emotion from either two of the three conditions matched is called 'shared'. The unexplained proportion of time was only 17.7% (bootstrapped 95% CI: 12.0% - 24.4%; dark gray bar in Fig. 4.3A). We found that the context-only condition was exclusively accurate 16.8% of the time (bootstrapped 95% CI: 10.2% - 24.8%; red bar in Fig. 4.3A). This was significantly more than the proportion of time when the blur-only condition exclusively matched (mean: 2.4%; bootstrapped 95% CI: 0.89% - 4.43%; blue bar in Fig. 4.3A; $p < 0.001$, permutation test). The proportion of time when the character-only condition exclusively matched the fully informed condition was 16.8% (bootstrapped 95% CI: 9.79% - 26.2%), the magnitude of which was comparable to that of the context-only condition ($p = 0.494$, permutation test) and was significantly larger than that explained only by the blur-only condition ($p < 0.001$, permutation test).

In accordance with the analysis in Chen and Whitney (2019), we also used linear regression models to estimate the degree to which variance in emotion tracking in the fully informed condition was explained only by the character, the context, or the blurred mask. We focused on the most selected emotion category at every time point and dummy coded the 11 emotion categories using 10 dichotomous variables. This variable transformation process was done for every condition separately. To estimate the proportion of unique variance explained by context, we first constructed a full model using the character-only variables, the context-only variables, and the blur-only variables to predict the fully informed emotion variables of the visible target. This linear full model performed well and explained a total of about 87.1% of the variance in emotion ratings (bootstrapped 95% CI: 81.8% – 92.0%). A second character-based model was created by using only the character-only variable and the blur-only variables to predict the fully informed variables 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. Similar procedures were carried out to estimate the unique variance of character-only variables and blur-only variables. The amount of variance that is explained by the full model but does not belong to either the context-only, the character-only, or the blur-only is considered shared variance among variables of two or more conditions. The proportion of unique variance in fully informed ratings that could only be explained by context-only ratings, but not character-only ratings or blur-only ratings, was 19.4% (bootstrapped 95% CI: 10.0% – 34.3%; Fig. 4.3C, red bar). This was significantly more than the unique variance explained by the blur-only variables (mean: 2.60%; bootstrapped 95% CI: 1.44%–3.95%; Fig. 4.3C, blue bar; $p < 0.001$, permutation test). The proportion of unique variance explained by the character-only condition was 19.2% (bootstrapped 95% CI: 11.7% – 30.0%; Fig. 4.3C, green bar), the magnitude of which was comparable to the unique variance explained by the context-only condition ($p = 0.510$, permutation test), and was significantly larger than that explained only by the blur-only condition ($p < 0.001$, permutation test). We confirmed that the unique contribution of context remained significant in the subset of video durations with incongruent emotions between character and context (Fig. C2 in *Appendix C*) and the subset with opposite valence between character and context (Fig. C3 in *Appendix C*). The contribution of context also remained significant in non-Hollywood movie clips (Fig. C4 in

Appendix C), videos without any other social agent or character present (Fig. C5 in *Appendix C*), and videos which participants reported to be not at all familiar with (Fig. C6 in *Appendix C*). Similarly, we did not find a significant difference in context effects attributable to participants' gender despite the imbalance in sample sizes of gender groups (Fig. C7 in *Appendix C*). Taken together, the results suggest that additional visual context information can shift perception of emotion from one category to another. Without the context, we would often misperceive a person's emotion over time.

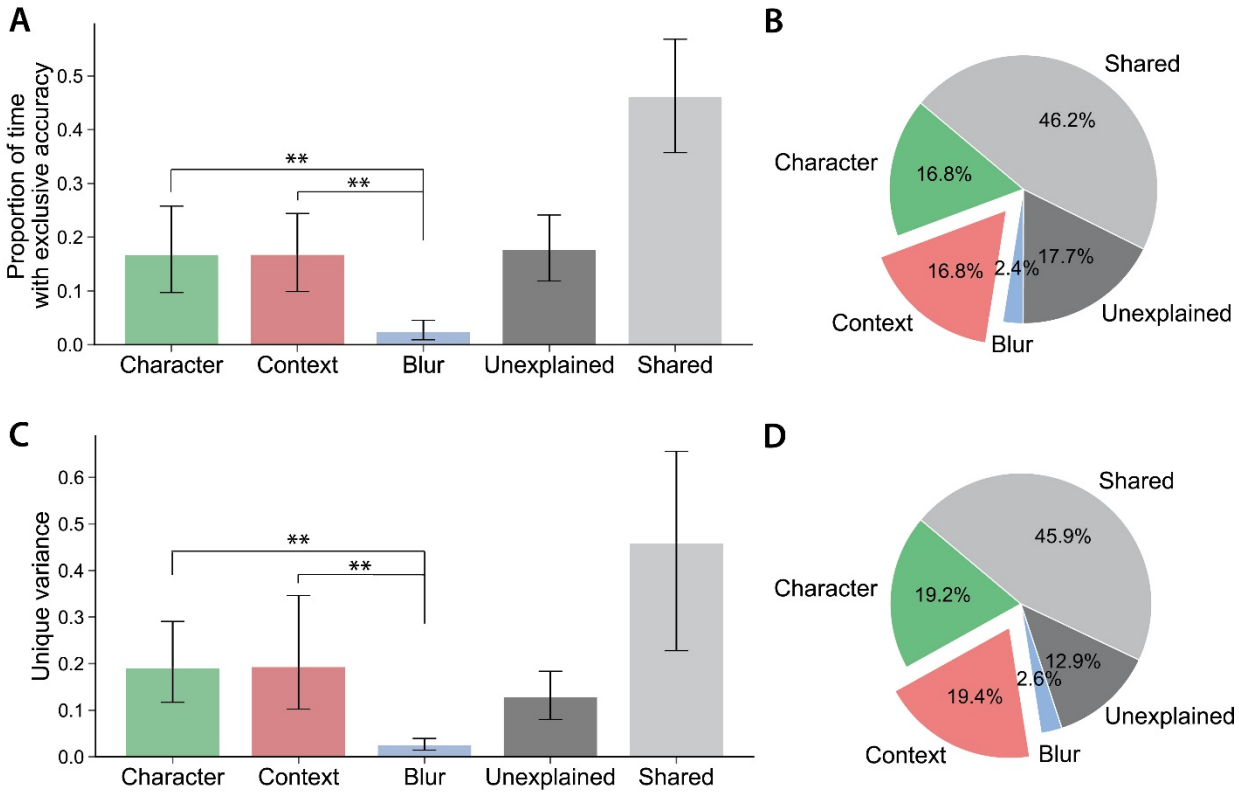


Fig. 4.3. (A-B) Proportion of time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. (C-D) Proportion of unique variance in the fully informed emotion ratings that could only be explained by character-only emotion ratings (in green), context-only emotion ratings (in red), and blur- only affect ratings (in blue). Light gray bar and pie show the proportion of variance shared between two or more than two types of ratings. The pie charts are redundant, but they show the cumulative variance sums to be one. Error bars represent bootstrapped 95% CI. * $p < 0.001$.

Discussion

Our results demonstrate that both visual scene context and character (face and body) information are essential to correctly interpret emotion categories. When face and body information was unavailable, observers could nevertheless infer emotion over time accurately, robustly, and with high agreement. Beyond the information available from face and body, visual scene context contributes a significant amount of unique information – as much as that from the

face and body. Background contextual information is therefore often necessary to most accurately recognize emotion category.

The results confirm and extend the findings of Chen & Whitney (2019), which demonstrate the essential contribution of visual context when tracking the affective dimensions of valence and arousal. Both affective dimensions and discrete emotion categories are important theoretical approaches to characterize emotion experience, and we show that visual context shapes the perception of emotion regardless of whether the emotion is reported as dimensional or categorical.

Our study corroborates the inferential tracking (IET) method and highlights its advantages in characterizing emotion when viewing ecologically valid and dynamic stimuli. This technique allows a large amount of data to be collected relatively quickly, as it leverages the rich variation in emotions as they unfold over time. The method also allows for straightforward descriptive and statistical estimations such as between-subject agreement and inferential tracking accuracy, which can be reliably derived from the data. Finally, the IET technique can be easily extended to test hypotheses in other domains of psychology, including research on cognitive and personality factors in emotion recognition and emotional intelligence.

Limitations and future directions. Although our results show that perception of emotion categories on average is heavily influenced by context, different emotion categories may be more or less susceptible to contextual influences by different degrees. A previous study found that fearful contexts may have a larger influence on neutral facial expressions than happy contexts do (Calbi et al., 2017), because fearful contexts may contain direct cues that elicit danger and activate defensive responses. The magnitude of contextual influences has also been linked to how stereotypical or diagnostic facial expressions are (Wieser & Brosch, 2012). For example, context may to be especially influential when facial expression is either ambiguous (e.g. surprised faces) or expression-less (e.g. neutral faces), because emotional information may be difficult to derive from facial features alone. Relatedly, it has also be suggested that the magnitude of contextual effects may be determined by how specific emotions can be confused with others (Aviezer et al., 2017; Aviezer et al., 2012). For example, an angry face might be more affected by a context indicating disgust, and less by a happy context, because the facial expressions of anger and disgust are perceptually similar and thus more confusable. A further analysis of our results shows that some specific emotions including fear, anger, happiness, and surprise have significant contextual effects ($p < 0.05/11$, after Bonferroni correction). The other emotion categories show trending effects of context, as well, but comparing directly across different emotion categories is not justified because the frequency of different emotion categories was not explicitly balanced in our study, and some emotions occurred far less frequently than others (Fig. 4.2B). Therefore, future studies that carefully balance the frequency of different emotions are needed to fairly compare the effects of context on specific emotion categories.

Likewise, future studies will be valuable in establishing that context-modulated categorical emotion perception occurs with lab-induced emotions. In the present experiments, we considered the group consensus of emotional interpretations under the fully informed condition as a useful and practical approximation of ground truth. The fully informed condition included all the visual information in the scene, so it is the closest to the default state observers encounter in typical circumstances. However, we do not know the actors' actual or intended emotions in our videos. Therefore, future studies can examine whether the context effects generalize to lab-induced emotions, which may be a more direct way to establish an alternative ground truth. Of

course, the ecological validity of lab-induced emotions may be compromised because the context and the lab setting may not approximate real-life situations as well as home videos or movie clips with professional actors.

It is also worth noting that the current study focused on how spatial context gives rise to the perception of emotion. Future studies are needed to investigate the temporal modulation of face, body, and contextual information. Different visual stimuli tend to change at different rates in the physical world (Stigliani et al., 2015). For example, visual context such as scenes are typically stationary and seem to vary at slow rates. In contrast, faces and bodies are dynamic and might change at much faster rates. However, it remains unknown whether face and visual context information manifest different temporal characteristics, whether they afford information at different timescales, and whether the visual system leverages the use of these sources of information at different temporal frequencies.

At a broader level, our findings are consistent with and extend a large body of vision research showing that visual contextual information actively interacts with local visual processing and directs perceptual interpretations (Albright & Stoner, 2002; Schwartz et al., 2007). Context strongly influences the perception of low-level visual features, such as brightness (Adelson, 2000), orientation (Gibson, 1937), motion (Wohlgemuth, 1911), and shadows (Rensink & Cavanagh, 2004). The visual system also implicitly and rapidly extracts contextual information from scenes to facilitate the recognition of individual objects (Bar, 2004; Biederman et al., 1982). For example, if observers have identified the context of a kitchen scene, they can infer that a fridge is probably present even without perceiving the fridge directly (Biederman et al., 1982). The perception of facial emotions is also influenced by the other faces presented nearby or in the past (Haberman & Whitney, 2007; Liberman et al., 2018; Mumenthaler & Sander, 2012). Our results extend previous work substantially, demonstrating that dynamic emotion perception is not just a product of facial expressions per se, but also incorporates non-face contextual information. The widespread evidence for contextual effects in vision and cognition suggests that the analysis and integration of context information is likely a fundamental process throughout the brain.

Integrating faces and context. Contrary to some seminal work (Ekman, 1992), the results here and in many previous papers show that perceiving emotion category is not simply an issue of registering facial expressions, per se (Aviezer et al., 2012; Barrett et al., 2019; Calbi et al., 2017; Kret & de Gelder, 2010). In fact, accumulating evidence suggests that emotion from facial expressions is inherently noisy, ambiguous, and uncertain (Hassin et al., 2013; Russell, 2016). So, how do observers combine image cues from different sources to estimate emotion, and what determines the importance the observer places on either facial or contextual cues? One optimal strategy to generate an accurate estimate of perceived emotion is to evaluate the trustworthiness of different sources of information and then place higher weights on cues that are less ambiguous and more reliable. Our data can address this. Because our tracking technique allows us to quantify facial ambiguity using between-subject agreement, we can evaluate whether background context is particularly useful in those cases where facial expressions are uncertain. We found a negative correlation between facial ambiguity and reliance on context: across all video clips, visual context had a significantly larger influence when facial expressions were more ambiguous (Fig. C8 in *Appendix C*).

Viewed in this light, the results seem consistent with the idea of emotion recognition as a kind of Bayesian inference of others' minds (Saxe & Houlihan, 2017). Emotions are internal

experiences and observers need to perform an inverse inference to use observed effects to infer underlying emotions. Formalized in this way, emotion inference requires the integration of signals of varying degrees of uncertainty. Of course, among these signals is facial expression, background context, and biological motion (e.g. Atkinson et al., 2004), as tested here, but there are a variety of other types of information and modalities like audition (Schirmer & Adolphs, 2017) and somatosensation (Kragel & LaBar, 2016) that could be combined. Priors, expectations, and rewards could also play a role (Ong et al., 2015; Ong et al., 2016). The present study, and the IET method in particular, could help pave the way toward quantitatively modeling the combination of different affective cues for the purpose of accurate emotion perception.

The IET method speaks directly to the overemphasis on facial expression in emotion and affect research. This narrow focus has pervaded the science of emotion for decades and has inadvertently led to the development of artificial intelligence systems in commercial and educational settings that analyze emotions based solely on facial expressions (e.g. Zeng et al., 2018; Affectiva.com; Microsoft Azure). Our findings reveal that without considering the context, AI systems will fall far short of fully understanding human emotion recognition and achieving genuine emotional intelligence. To overcome this setback, it is important for the scientific study of emotion to devote more to capturing the rich and distinctive landscape of emotion in context.

Chapter 5: Conclusions

As we have discussed, there has been a dominant view in emotion and affect research that emotion recognition is largely based on the ability to read facial expressions. In this set of experiments, we have demonstrated that dynamic emotion perception is much more than simply registering facial expressions: it also requires a significant integration of non-face contextual information. These results bring us closer to answering how the brain achieves emotion recognition with remarkably speed and efficiency.

In Chapter 2, we found that even when no face information was present, the visual context was sufficient to infer valence and arousal over time, and different observers agreed about the affective information provided by the context. We further demonstrated that participants used unique information about the background context, independent of any face information, to accurately register and track affect. 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. These results suggest that context is an essential component of emotion recognition.

In Chapter 3, we provided the first measure of the speed of context-based dynamic emotion perception and showed that the context is processed with a remarkably short latency. Using cross-correlation analysis, we showed that inferring emotion from only contextual information is essentially as fast as using all available information including facial expressions. With empirical experimental manipulations, we further demonstrated the precision of our method by showing that it could resolve a 100 msec temporal lag. Seemingly complex context-dependent emotional inference and recognition is far more efficient than previously assumed.

In Chapter 4, we confirmed and extended the findings of Chapter 2 by showing that visual context remains contributing uniquely and as much as face and body regardless of whether the emotion is reported as dimensional or categorical. Without the context, we would often misperceive a person's emotion over time.

In sum, the experiments address the original hypothesis that context could be a primary cue to emotion, not simply a secondary or modulatory cue. Our research reveals that context is both sufficient and necessary to accurately recognize emotion, and the available information from context is processed rapidly. The role of context in emotion recognition cannot be understated - emotion recognition is, at its heart, an issue of context as much as it is about faces.

References

Adelson, E. H. (2000). 24 Lightness Perception and Lightness Illusions. In M. Gazzaniga (Eds.), *The New Cognitive Neurosciences* (2nd Ed., pp. 339-351). Cambridge, MA: MIT Press.

Albright, T. D., & Stoner, G. R. (2002). Contextual influences on visual processing. *Annual Review of Neuroscience*, *25*(1), 339–379. https://doi.org/10.1146/annurev.neuro.25.112701.142900

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, *33*(6), 717–746. https://doi.org/10.1068/p5096

Aviezer, H., Bentin, S., Dudarev, V., & Hassin, R. R. (2011). The automaticity of emotional face-context integration. *Emotion*, *11*(6), 1406–1414. https://doi.org/10.1037/a0023578

Aviezer, H., Ensenberg, N., & Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, *17*, 47–54. https://doi.org/10.1016/j.copsyc.2017.06.006

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, *338*(6111), 1225–1229. https://doi.org/10.1126/science.1224313

Bahrami, W., Rangin, H., & Rangin, K. (2010). A Two-parameter Generalized Skew-Cauchy Distribution. *Journal of Statistical Research of Iran JSRI*, *7*(1), 61-72.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Barrett, L. F., & Kensinger, E. A. (2010). Context Is Routinely Encoded During Emotion Perception. *Psychological Science*, *21*(4), 595–599. https://doi.org/10.1177/0956797610363547

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in Emotion Perception. *Current Directions in Psychological Science*, *20*(5), 286–290. https://doi.org/10.1177/0963721411422522

Bayazit, M., & Önöz, B. (2007). To prewhiten or not to prewhiten in trend analysis? *Hydrological Sciences Journal*, *52*(4), 611–624. https://doi.org/10.1623/hysj.52.4.611

Betz, N., Hoemann, K., & Barrett, L. F. (2019). Words are a context for mental inference. *Emotion, 19*(8), 1463–1477. doi:10.1037/emo0000510

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-x

Bradley, M. M., & Lang, P. J. (1999) *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Tech. Rep. C-1). The Center for Research in Psychophysiology. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi_10.1.1.306.3881&rep_rep1&type_pdf

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://www.ncbi.nlm.nih.gov/pubmed/9176952

Calbi, M., Heimann, K., Barratt, D., Siri, F., Umiltà, M. A., & Gallese, V. (2017). How Context Influences Our Perception of Emotional Faces: A Behavioral Study on the Kuleshov Effect. *Frontiers in Psychology*, *8*, 1684. https://doi.org/10.3389/fpsyg.2017.01684

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641–651. https://doi.org/10.1038/nrn1724

Calder, A. J., Ewbank, M., & Passamonti, L. (2011). Personality influences the neural responses to viewing facial expressions of emotion. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1571), 1684–1701. https://doi.org/10.1098/rstb.2010.0362

Chen, Z., & Whitney, D. (2019). Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, *116*(15), 7559–7564. doi: 10.1073/pnas.1812250116

Chen, Z., & Whitney, D. (2020). Inferential emotion tracking (IET) reveals the critical role of context in emotion recognition. *Manuscript in press in Emotion.*

Clore, G. L., & Ortony, A. (2013). Psychological Construction in the OCC Model of Emotion. *Emotion Review: Journal of the International Society for Research on Emotion*, *5*(4), 335–343. https://doi.org/10.1177/1754073913489751

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. San Diego: Academic press.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75–93. https://doi.org/10.1037/emo0000302

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900-E7909. doi: 10.1073/pnas.1702247114

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369–382. https://doi.org/10.1038/s41562-019-0533-6

Dalili, M. N., Penton-Voak, I. S., Harmer, C. J., & Munafò, M. R. (2015). Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychological Medicine*, *45*(06), 1135–1144. https://doi.org/10.1017/S0033291714002591

de Gelder, B., & Van den Stock, J. (2011). Real faces, real emotions: perceiving facial expressions in naturalistic contexts of voices, bodies and scenes. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford Handbook of Face Recognition* (pp. 535-550). New York: Oxford University Press.

de Gelder, B., De Borst, A. W., & Watson, R. (2015). The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2), 149-158.

de Gelder, B., Meeren, H. K. M., Righart, R., van den Stock, J., van de Riet, W. A. C., & Tamietto, M. (2006). Beyond the face: exploring rapid influences of context on face processing. *Progress in Brain Research*, *155*, 37–48. https://doi.org/10.1016/S0079-6123(06)55003-4

Dean, R. T., & Dunsmuir, W. T. M. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods*, *48*(2), 783–802. https://doi.org/10.3758/s13428-015-0611-2

Del Sole, A. (2018). Introducing Microsoft cognitive services. In *Microsoft Computer Vision APIs Distilled* (pp. 1-4). Apress, Berkeley, CA.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (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.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. https://www.ncbi.nlm.nih.gov/pubmed/5811531

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Falagiarda, F., & Collignon, O. (2019). Time-resolved discrimination of audio-visual emotion expressions. *Cortex*, *119*, 184–194. https://doi.org/10.1016/j.cortex.2019.04.017

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*(1), 153–166. https://doi.org/10.1037/0022-3514.69.1.153

Fernández-Dols, J. M., & Crivelli, C. (2013). Emotion and expression: Naturalistic studies. *Emotion Review: Journal of the International Society for Research on Emotion*, *5*(1), 24–29. https://doi.org/10.1177/1754073912457229

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, *106*(3), 1389-1398. https://doi.org/10.1152/jn.00904.2010

Fredrickson, B. L. (1998). What Good Are Positive Emotions? *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, *2*(3), 300–319. https://doi.org/10.1037/1089-2680.2.3.300

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gallese, V., & Sinigaglia, C. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences*, *15*(11), 512–519. https://www.sciencedirect.com/science/article/pii/S136466131100194X

Gibson, J. J. (1937). Adaptation, after-effect, and contrast in the perception of tilted lines. II. Simultaneous contrast and the areal restriction of the after-effect. *Journal of Experimental Psychology*, *20*(6), 553)–569.https://doi.org/10.1037/h0057585

Gopnik A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirchfield, & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge Univ Press, New York.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753. https://doi.org/10.1016/j.cub.2007.06.039

Harms, M. B., Martin, A., & Wallace, G. L. (2010). Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review*, *20*(3), 290–322. https://doi.org/10.1007/s11065-010-9138-6

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently Ambiguous: Facial Expressions of Emotions, in Context. *Emotion Review: Journal of the International Society for Research on Emotion*, *5*(1), 60–65. https://doi.org/10.1177/1754073912451331

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175

Jeffreys, H. (1961). *Theory of probability*, 3rd Edn. Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kayyal, M., Widen, S., & Russell, J. a. (2015). Context is more powerful than we think: Contextual cues override facial cues even for valence. *Emotion*, *15*(3), 287–291. https://doi.org/10.1037/emo0000032

Keltner, D., & Cordaro, D. T. (2015) Understanding multimodal emotional expressions: Recent advances in Basic Emotion Theory. In J. M. Fernández-Dols & J. A. Russell (Eds.), *The science of facial expression* (pp. 57–75). New York, NY: Oxford University Press.

Keltner, D., & Haidt, J. (1999). Social Functions of Emotions at Four Levels of Analysis. *Cognition & Emotion*, *13*(5), 505–521. https://doi.org/10.1080/026999399379168

Keltner, D., & Lerner, J. S. (2010). Emotion. In S. T. Fiske (Ed.), *Handbook of social psychology, Vol* (Vol. 1, pp. 317–352). John Wiley & Sons Inc, xv. https://psycnet.apa.org/fulltext/2010-03505-009.pdf

Kimchi, R., & Peterson, M. A. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, *19*(7), 660–668. https://doi.org/10.1111/j.1467-9280.2008.02140.x

Kimura, M., Kondo, H., Ohira, H., & Schröger, E. (2012). Unintentional Temporal Context–Based Prediction of Emotional Faces: An Electrophysiological Study. *Cerebral Cortex*, *22*(8), 1774–1785. https://doi.org/10.1093/cercor/bhr244

Kohler, C. G., Turner, T. H., Gur, R. E., & Gur, R. C. (2004). Recognition of facial emotions in neuropsychiatric disorders. *CNS Spectrums*, *9*(4), 267–274. https://doi.org/10.1017/s1092852900009202

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2010). Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophrenia Bulletin*, *36*(5), 1009–1019. https://doi.org/10.1093/schbul/sbn192

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, *65*, 23–36. https://doi.org/10.1016/j.imavis.2017.02.001

Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017). Emotion Recognition in Context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (cvpr)*, 1667–1675. https://doi.org/10.1109/CVPR.2017.212

Kragel, P. A., & LaBar, K. S. (2016). Decoding the Nature of Emotion in the Brain. *Trends in Cognitive Sciences*, *20*(6), 444–455. doi: 10.1016/j.tics.2016.03.011

Kret, M. E., & de Gelder, B. (2010). Social context influences recognition of bodily expressions. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *203*(1), 169–180. https://doi.org/10.1007/s00221-010-2220-8

Kret, M. E., Roelofs, K., Stekelenburg, J. J., & de Gelder, B. (2013). Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. *Frontiers in Human Neuroscience*, *7*, 810. https://doi.org/10.3389/fnhum.2013.00810

Kwiatkowski, D., Phillips, P. C. B., 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? *Journal of Econometrics*, *54*(1), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (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.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Press.

Liberman, A., Manassi, M., & Whitney, D. (2018). Serial dependence promotes the stability of perceived emotional expression depending on face similarity. *Attention, Perception & Psychophysics*, *80*(6), 1461–1473. https://doi.org/10.3758/s13414-018-1533-8

Liberman, A., Manassi, M., & Whitney, D. (2018). Serial dependence promotes the stability of perceived emotional expression depending on face similarity. *Attention, Perception & Psychophysics*, *80*(6), 1461–1473. https://doi.org/10.3758/s13414-018-1533-8

Liedtke, C., Kohl, W., Kret, M. E., & Koelkebeck, K. (2018). Emotion recognition from faces with in- and out-group features in patients with depression. *Journal of Affective Disorders*, 227, 817-823. https://doi.org/10.1016/j.jad.2017.11.085

Marsh, B. (2002). Heuristics as social tools. *New Ideas in Psychology*, *20*(1), 49–57. https://doi.org/10.1016/S0732-118X(01)00012-5

Masuda, T., Ellsworth, P. C., Mesquita, B., Leu, J., Tanida, S., & Van de Veerdonk, E. (2008). Placing the face in context: cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology*, *94*(3), 365–381. https://doi.org/10.1037/0022-3514.94.3.365

Matsumoto, D., Keltner, D., Shiota, M. N., O'Sullivan, M., & Frank, M. (2008). Facial expressions of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *The handbook of emotion* (pp. 211-234). New York: Guilford. Retrieved from https://psycnet.apa.org/fulltext/2008-07784-013.pdf

Matsumoto, D., Le Roux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., Yrizarry, N., Loewinger, S., Uchida, H., Yee, A., Amo, L., & Goh, A. (2000). A New Test to Measure Emotion Recognition Ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior*, *24*(3), 179-209.

Mattingley, J. B., Davis, G., & Driver, J. (1997). Preattentive filling-in of visual surfaces in parietal extinction. *Science*, *275*(5300), 671–674. https://doi.org/10.1126/science.275.5300.671

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. https://www.sciencedirect.com/science/article/pii/S1053811915008873

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human Abilities: Emotional Intelligence. *Annual Review of Psychology*, *59*(1), 507–536. https://doi.org/10.1146/annurev.psych.59.103006.093646

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion*, *3*(1), 97–105. https://doi.org/10.1037/1528-3542.3.1.97

Meeren, H. K. M., van Heijnsbergen, C. C. R. J., & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16518–16523. https://doi.org/10.1073/pnas.0507650102

Melloni, M., Lopez, V., & Ibanez, A. (2014). Empathy and contextual social cognition. *Cognitive, Affective & Behavioral Neuroscience*, *14*(1), 407–425. https://doi.org/10.3758/s13415-013-0205-3

Mumenthaler, C., & Sander, D. (2012). Social appraisal influences recognition of emotions. *Journal of Personality and Social Psychology*, *102*(6), 1118–1135. https://doi.org/10.1037/a0026885

Mumenthaler, C., & Sander, D. (2015). Automatic integration of social information in emotion recognition. *Journal of Experimental Psychology. General*, *144*(2), 392–399. https://doi.org/10.1037/xge0000059

Olsson, A., & Ochsner, K. N. (2008). The role of social cognition in emotion. *Trends in Cognitive Sciences*, *12*(2), 65–71. https://doi.org/10.1016/j.tics.2007.11.010

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162. https://doi.org/10.1016/j.cognition.2015.06.010

Ong, D., Asaba, M., & Gweon, H. (2016). Young children and adults integrate past expectations and current outcomes to reason about others' emotions. *CogSci*, 135–140. http://mindmodeling.org/cogsci2016/papers/0036/paper0036.pdf

Otten, M., Seth, A. K., & Pinto, Y. (2017). A social Bayesian brain: How social knowledge can shape visual perception. *Brain and Cognition*, *112*, 69–77. https://doi.org/10.1016/j.bandc.2016.05.002

Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: how much is enough? *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 230–242. https://doi.org/10.3758/CABN.10.2.230

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://www.ncbi.nlm.nih.gov/pubmed/9176953

Poncet, F., Baudouin, J.-Y., Dzhelyova, M. P., Rossion, B., & Leleu, A. (2019). Rapid and automatic discrimination between facial expressions in the human brain. *Neuropsychologia*, *129*, 47–55. https://doi.org/10.1016/j.neuropsychologia.2019.03.006

Probst, W. N., Stelzenmüller, V., & Fock, H. O. (2012). Using cross-correlations to assess the relationship between time-lagged pressure and state indicators: an exemplary analysis of North Sea fish population indicators. *ICES Journal of Marine Science: Journal Du Conseil*, 69(4), 670–681. https://doi.org/10.1093/icesjms/fss015

Razavi, S., & Vogel, R. (2018). Prewhitening of hydroclimatic time series? Implications for inferred change and variability across time scales. *Journal of Hydrology*, *557*, 109–115. https://doi.org/10.1016/j.jhydrol.2017.11.053

Rensink, R. A., & Cavanagh, P. (2004). The influence of cast shadows on visual search. *Perception*, *33*(11), 1339–1358. https://doi.org/10.1068/p5322

Reschke, P. J., Walle, E. A., Knothe, J. M., & Lopez, L. D. (2019). The influence of context on distinct facial expressions of disgust. *Emotion*, *19*(2), 365–370. doi: 10.1037/emo0000445

Righart, R., & de Gelder, B. (2006). Context influences early perceptual analysis of faces--an electrophysiological study. *Cerebral Cortex*, *16*(9), 1249–1257. https://doi.org/10.1093/cercor/bhj066

Righart, R., & de Gelder, B. (2008a). Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, Affective & Behavioral Neuroscience*, *8*(3), 264–272. https://doi.org/10.3758/CABN.8.3.264

Righart, R., & De Gelder, B. (2008b). Rapid influence of emotional scenes on encoding of facial expressions: an ERP study. *Social Cognitive and Affective Neuroscience*, *3*(3), 270-278.

Rozin, P., & Cohen, A. B. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, *3*(1), 68–75. https://doi.org/10.1037/1528-3542.3.1.68

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A. (2016). A sceptical look at faces as emotion signals. In C. Abell, & J. Smith (Eds.), *The Expression of Emotion: Philosophical, Psychological and Legal Perspectives* (pp. 157–172). Cambridge University Press Cambridge. https://doi.org/10.1017/cbo9781316275672.008

Russell, J. A., & Dols, J. M. F. (1997). *The Psychology of Facial Expression (Vol. 131)*. Cambridge university press Cambridge.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15–21. https://doi.org/10.1016/j.copsyc.2017.04.019

Schirmer, A., & Adolphs, R. (2017). Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence. *Trends in Cognitive Sciences*, *21*(3), 216–228. https://doi.org/10.1016/j.tics.2017.01.001

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, *8*(7), 522–535. https://doi.org/10.1038/nrn2155

Shiota, M. N., Campos, B., Oveis, C., Hertenstein, M. J., Simon-Thomas, E., & Keltner, D. (2017). Beyond happiness: Building a science of discrete positive emotions. *The American Psychologist*, *72*(7), 617–643. https://doi.org/10.1037/a0040456

Shumway, R. H., & Stoffer, D. S. (2014). *Time Series Analysis and Its Applications*. Springer, New York. https://play.google.com/store/books/details?id=N-EYswEACAAJ

Sims, C. A. (1988). Bayesian skepticism on unit root econometrics. *Journal of Economic Dynamics & Control*, *12*(2), 463–474. https://doi.org/10.1016/0165-1889(88)90050-4

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(48), 15997–16008. https://doi.org/10.1523/JNEUROSCI.1676-14.2014

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(36), 12412–12424. https://doi.org/10.1523/JNEUROSCI.4822-14.2015

Ventura-Bort, C., Löw, A., Wendt, J., Dolcos, F., Hamm, A. O., & Weymar, M. (2016). When neutral turns significant: brain dynamics of rapidly formed associations between neutral stimuli and emotional contexts. *The European Journal of Neuroscience*, *44*(5), 2176–2183. https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13319

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, *23*(3), 230–249. https://doi.org/10.1016/0022-1031(87)90034-5

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*(6), 807–814. https://doi.org/10.1037/emo0000185

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Wieser, M. J., & Brosch, T. (2012). Faces in context: a review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, *3*, 471. https://doi.org/10.3389/fpsyg.2012.00471

Wohlgemuth, A. (1911). On the after-effect of seen movement. *British Journal of Psychology Monograph Supplements*, *1*, 1–117.

Yamanashi Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications*, *7*, 1–10. https://doi.org/10.1038/ncomms13186

Yang, Y.-H., & Yeh, S.-L. (2018). Unconscious processing of facial expression as revealed by affective priming under continuous flash suppression. *Psychonomic Bulletin & Review*, *25*(6), 2215–2223. https://doi.org/10.3758/s13423-018-1437-6

Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T.-C., & Qu, H. (2020). EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos. *IEEE Transactions on Visualization and Computer Graphics, 1*, 1-1. https://doi.org/10.1109/TVCG.2019.2963659

Zhaoping, Li. (2000). Pre-attentive segmentation in the primary visual cortex. *Spatial Vision*, *13*(1), 25–50. https://doi.org/10.1163/156856800741009

Zhou, H., Majka, E. A., & Epley, N. (2017). Inferring Perspective Versus Getting Perspective: Underestimating the Value of Being in Another Person's Shoes. *Psychological Science*, *28*(4), 482–493. https://doi.org/10.1177/0956797616687124

Appendix A: Supplemental Figures for Chapter 2

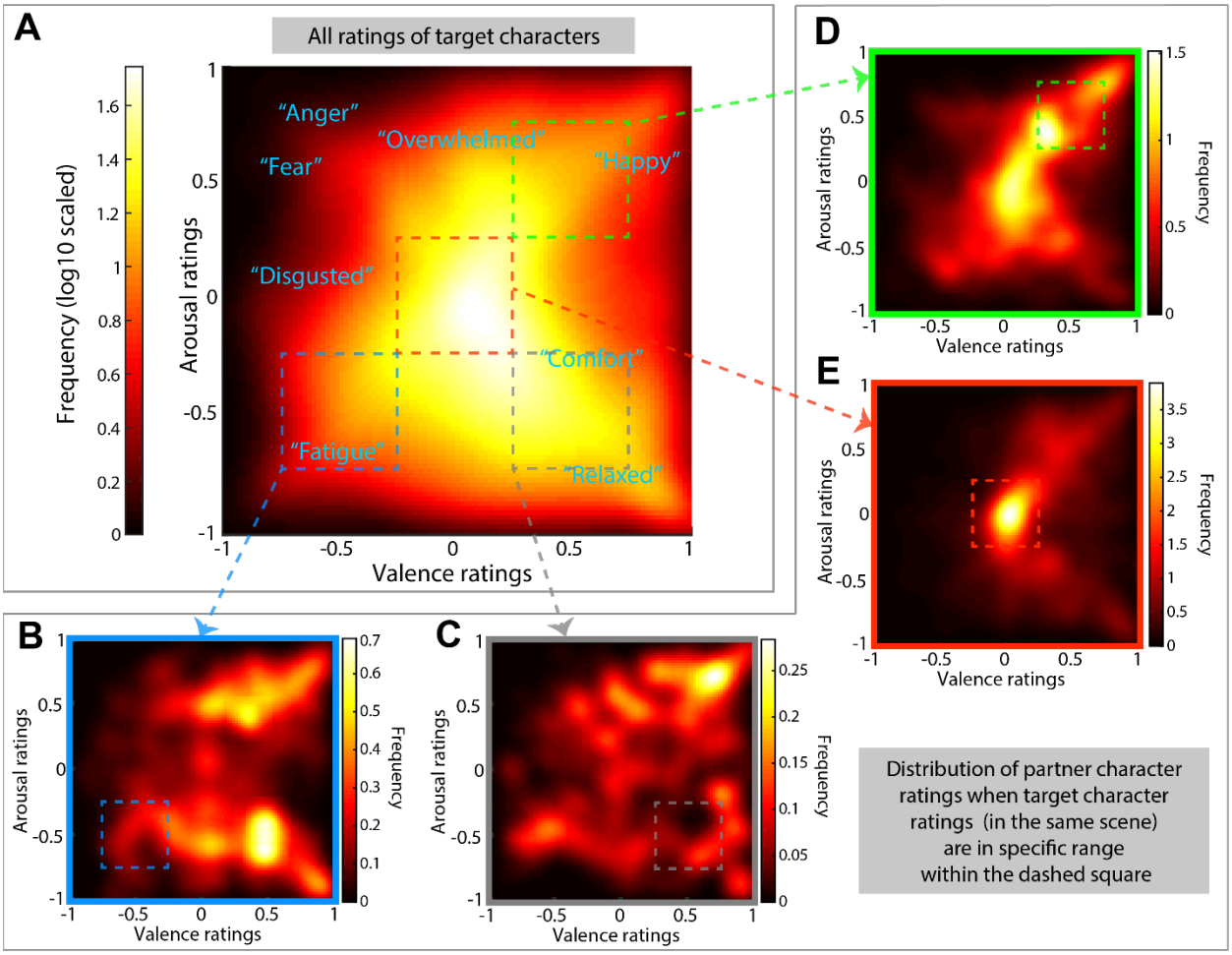


Fig. A1. (A) Distribution of raw fully-informed ratings across the valence-arousal space (data collapsed across experiment 1 and 2). Colors represent the log 10-transformed value of counts within each interval. Affective words have been placed at locations within this space, representing their approximate 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 does 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. 2.2). 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.

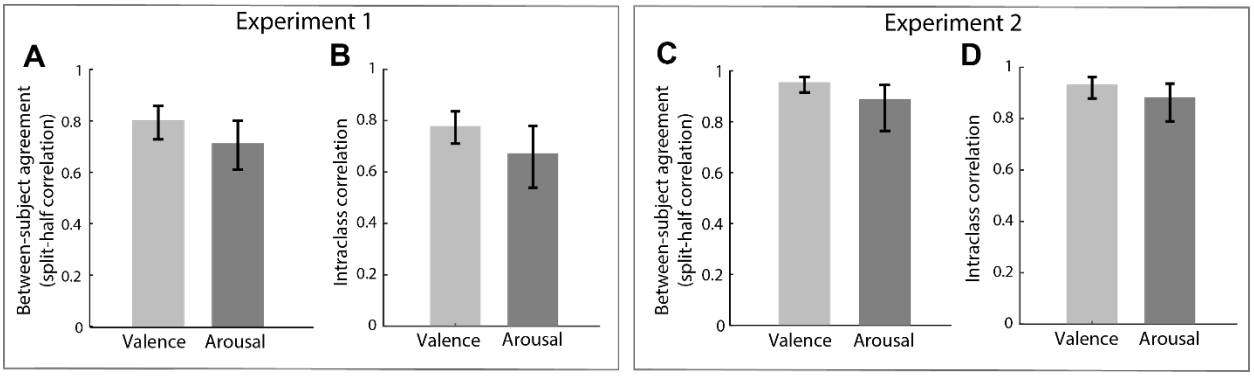


Fig. A2. Between-subject agreement of inferred ratings (when characters were masked and invisible) evaluated by split-half correlation and intraclass correlation. (A & C) In split-half correlation, inferred affect ratings provided by different subjects on each video clip were split into two halves, and 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).

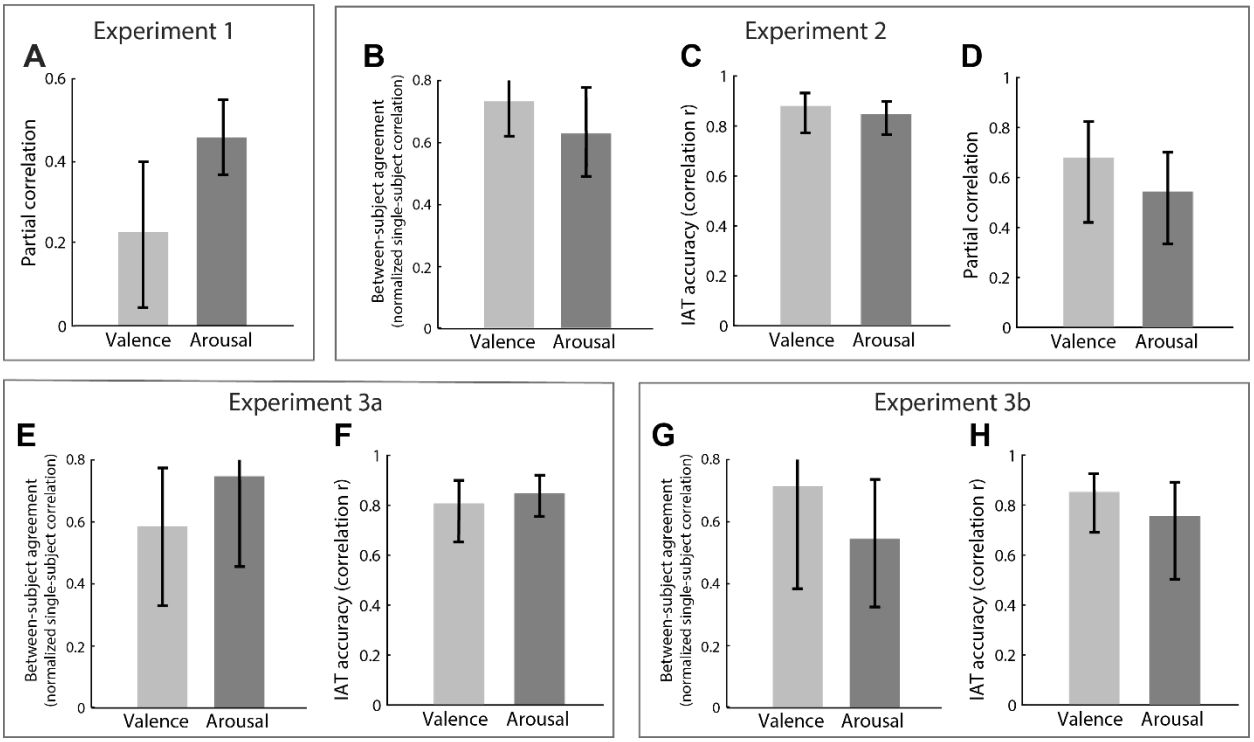


Fig. A3. (A) Mean partial correlations (separately for valence and arousal) between inferred affect ratings of the invisible target and fully-informed affect ratings of the visible target, when controlling for fully-informed affect ratings of the visible partner. (B, E, G) Between-subject agreement of 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).

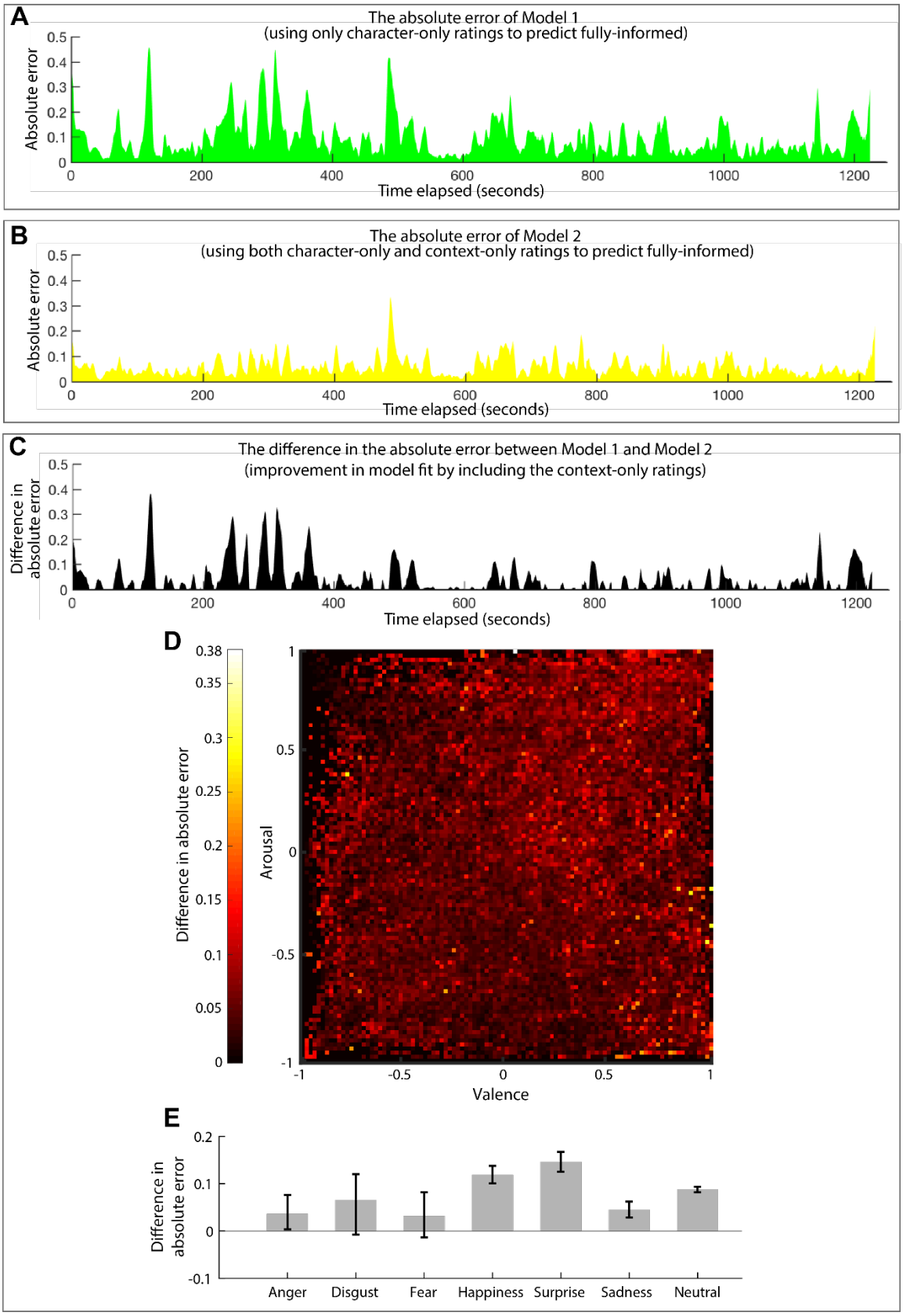

A
The absolute error of Model 1
(using only character-only ratings to predict fully-informed)
Absolute error
Time elapsed (seconds)
B
The absolute error of Model 2
(using both character-only and context-only ratings to predict fully-informed)
Absolute error
Time elapsed (seconds)
C
The difference in 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. A4. 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 (Del Sole, 2018), 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.

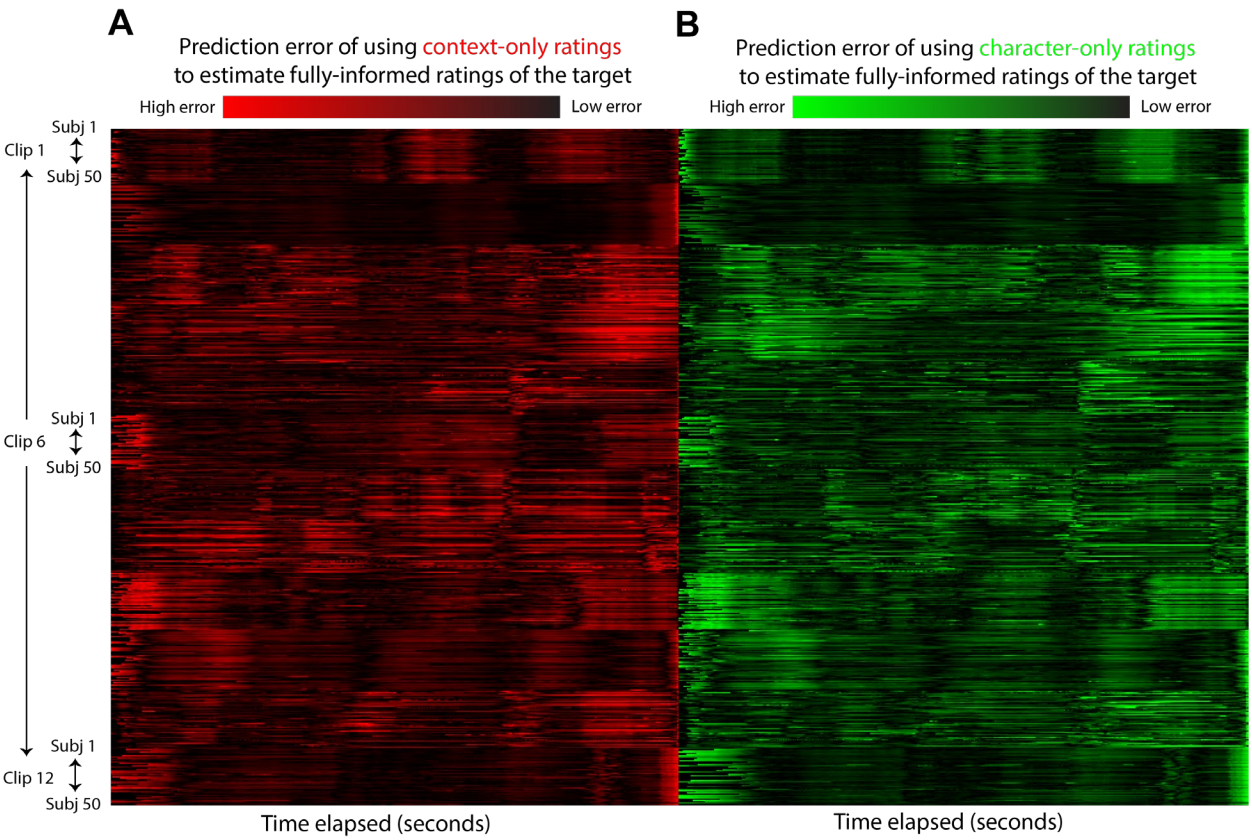


Fig. A5. Affect perceived from only contextual information is nearly as accurate as affect perceived from facial and bodily information. (A) The prediction (absolute) error of using context-only ratings to estimate fully-informed affect ratings of the target characters (indicated by the intensity of 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.

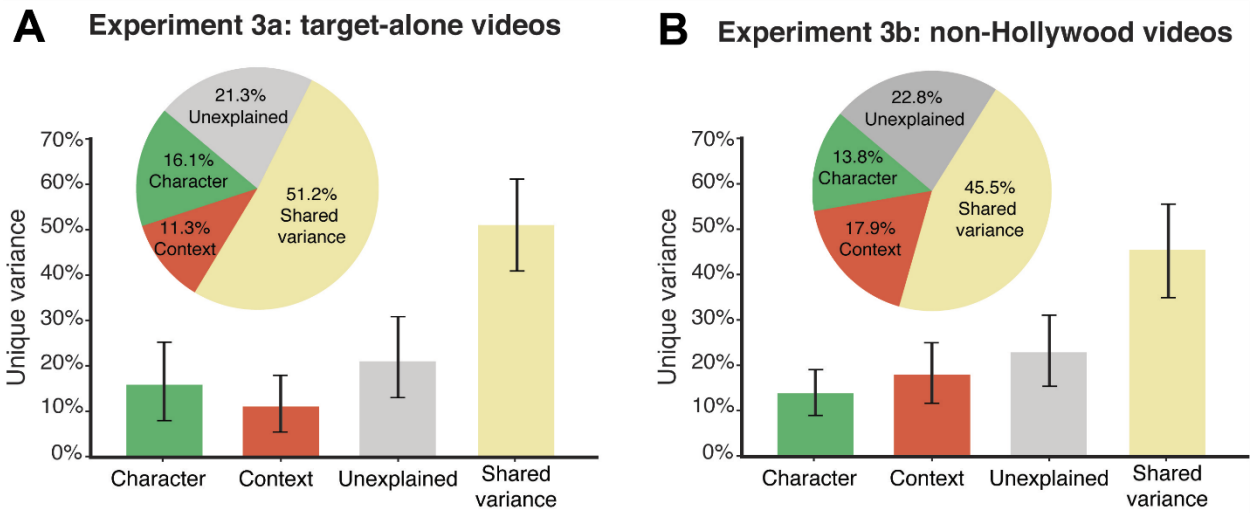


Fig. A6. Proportion of unique variance in the fully-informed affect ratings that could only be explained by context-only affect ratings (in red) versus character-only affect ratings (in green). Yellow bar and pie show the proportion of variance shared between context-only and character-only ratings. (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)

Appendix B: Supplemental Figures for Chapter 3

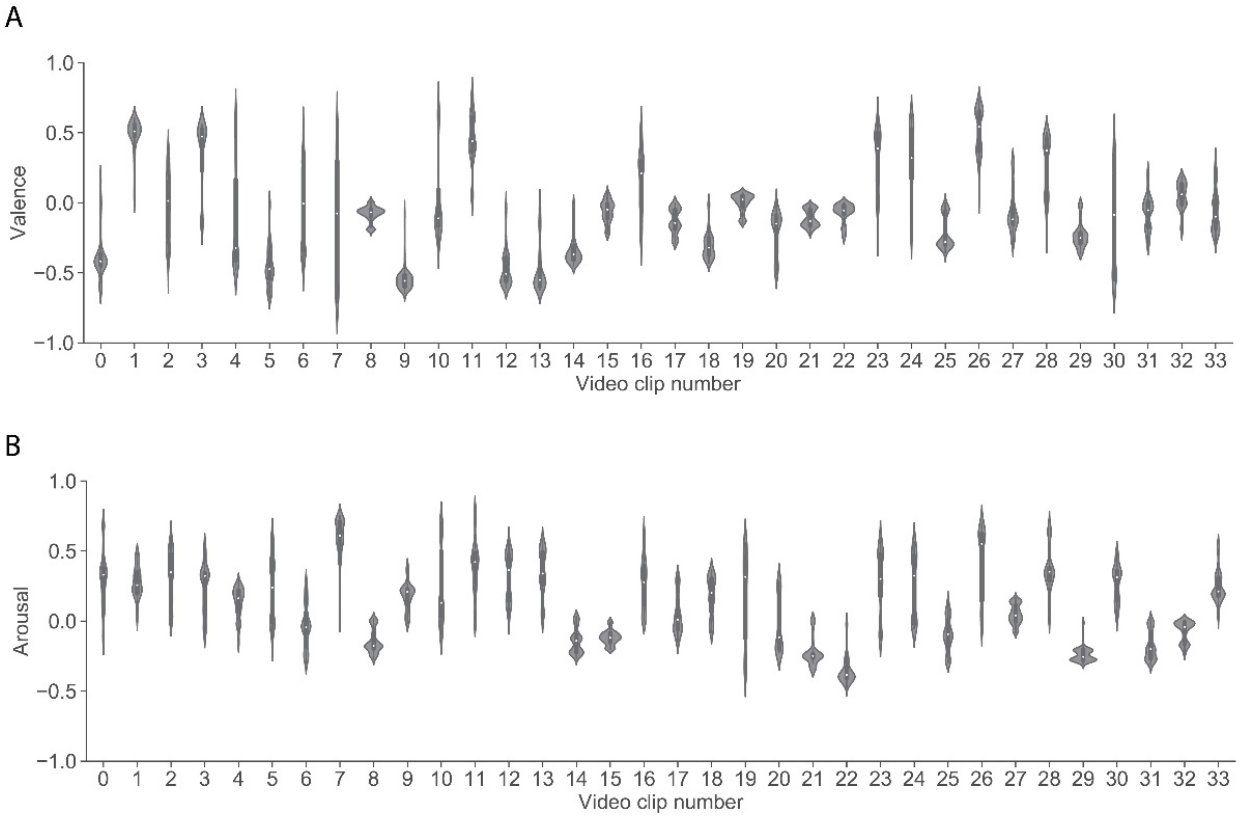


Fig. B1. The distribution of valence (A) and arousal (B) ratings in all 34 video clips. To effectively evaluate tracking performance, we used video clips that contained variations in emotion. This means that many videos in our dataset contained more than 1 emotion and some transitions from positive to negative emotions or from negative to positive. To quantify this, we averaged the valence/arousal ratings of the target character within every 1-second bin along the time course of every video. The violin plots show the distribution of valence and arousal for all binned time points within a given video (34 videos in total). In most of the videos, the affect was quite heterogeneous.

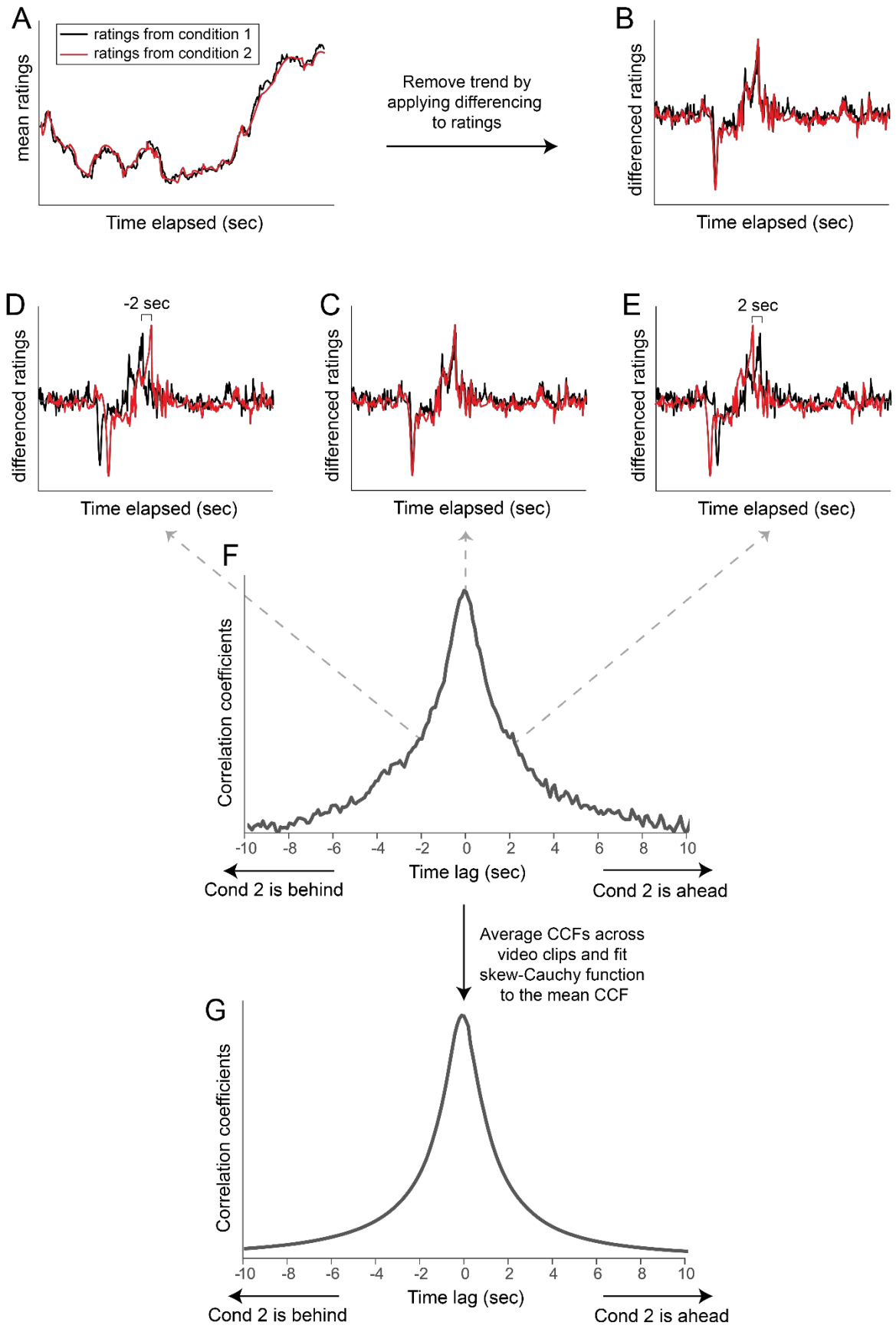

A
ratings from condition 1
ratings from condition 2
mean ratings
Time elapsed (sec)
Remove trend by applying differencing to ratings
B
differenced ratings
Time elapsed (sec)
D
-2 sec
differenced ratings
Time elapsed (sec)
C
differenced ratings
Time elapsed (sec)
E
2 sec
differenced ratings
Time elapsed (sec)
F
Correlation coefficients
-10 -8 -6 -4 -2 0 2 4 6 8 10
Time lag (sec)
Cond 2 is behind
Cond 2 is ahead
Average CCFs across video clips and fit skew-Cauchy function to the mean CCF
G
Correlation coefficients
-10 -8 -6 -4 -2 0 2 4 6 8 10
Time lag (sec)
Cond 2 is behind
Cond 2 is ahead

Fig. B2. Calculating a cross correlation function (CCF) to measure lag in emotion tracking. (A) We obtained mean affect ratings for each individual video in each condition by averaging across responses from all participants. The visualizations here are for one single video; the same analysis was performed all 34 videos. (B) We then transformed the data for each video to make it stationary by applying differencing (Sims, 1988), which involves subtracting every value x_t from x_{t+1} to obtain successive differences between adjacent values in time. As an alternative, establishing stationarity with a pre-whitening approach (Shumway & Stoffer, 2011) did not change the results. The validity of the approach is confirmed by the near flat and zero CCF functions in the permuted null distributions (see Fig. 3.2A). (C, D, E, F) We computed the correlation coefficient between ratings from both conditions after shifting one series of ratings with different time displacement relative to the other. (e.g. -2 sec for D, 0 sec for C, 2 sec for E). This was performed for all possible time displacements to obtain the continuous cross correlation function in F for every video. (G) The CCFs for individual videos are averaged to obtain the mean CCF across videos. A skew-Cauchy distribution (Bahrami et al., 2010) was fit to the mean CCF in order to capture the shape of the CCF and the time lag that has the highest correlation.

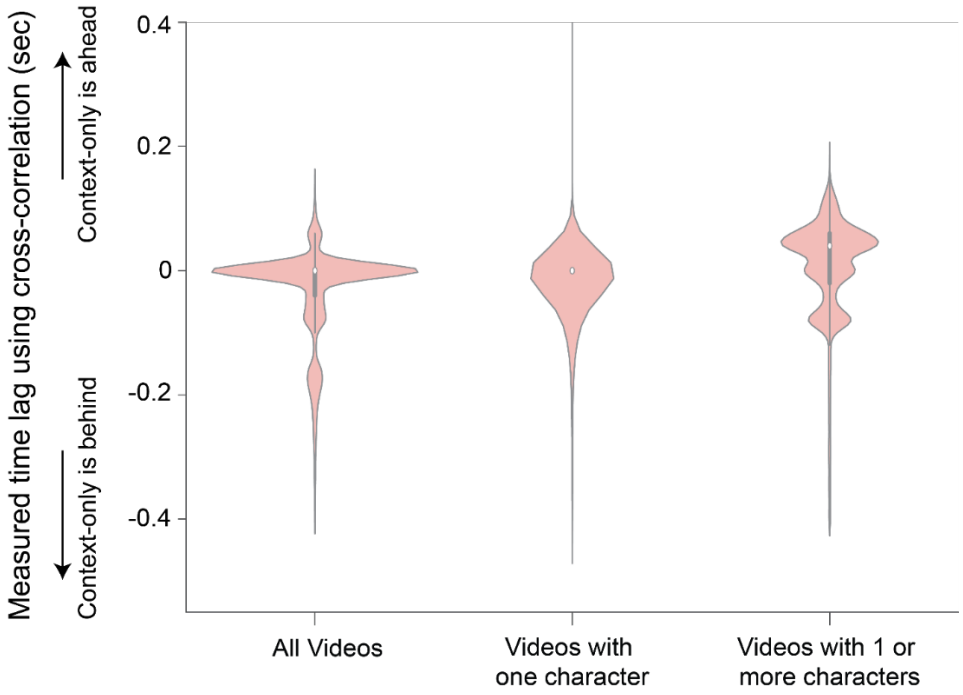


Fig. B3. Violin plots of the measured time lag in the context-only condition (experiment 1) using all videos (n = 34), videos with only one character (n = 9), and videos with one or more characters (n = 25). We split the video clips into two groups depending on whether there was a partner character shown in the context, and we quantified the measured peak lags in that subset of videos. We found that in both sets of the videos, the distribution of measured time lags was near zero and it did not show a clear trend for substantial lags.

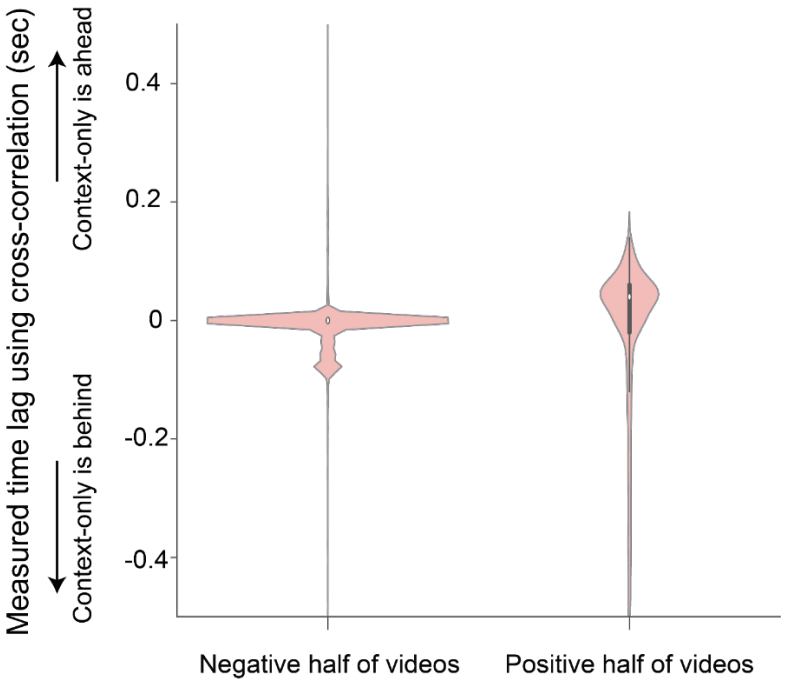


Fig. B4. The distribution of measured time lags in the context-only condition (experiment 1) for the subsets of videos with relatively negative or positive valence. We split the video clips into two halves depending on whether affect ratings of the target characters were, on average, more negative or positive. We then quantified the measured time lags in these two groups of videos. We found that in both halves of the videos, the distribution of measured time lags was near zero.

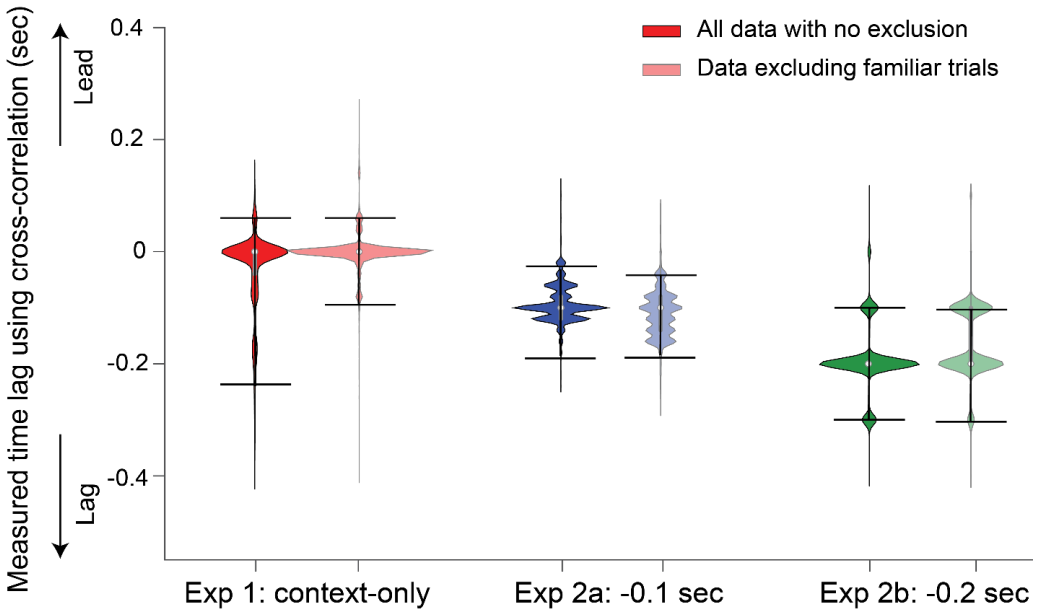


Fig. B5. The distribution of measured time lags for all experiments. On average, in 87% of all trials, participants reported that they had not seen the video content before participating in the experiments. We analyzed the data excluding trials in which participants reported familiarity. The distribution of measured time lags is similar to results obtained with data with no exclusion.

Appendix C: Supplemental Figures for Chapter 4

Fig. C1. Relative occurrence frequency of the most selected emotion category in all time points across all video stimuli for the fully informed condition, the context-only condition, and the character-only condition. Removing face and body information in the context-only condition does not dramatically shift the distribution of emotion categories chosen compared to that in the fully informed condition. The relative frequency of the chosen emotion categories here closely correlates with the frequency of emotion category words found in English (Fig. 4.2).

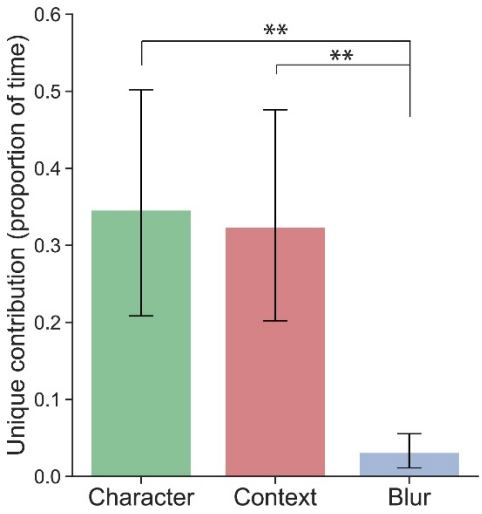


Fig. C2. Unique contribution of context versus character in the subset of video durations when the emotion implied by the context is incongruent with the emotion implied by the character. To quantify incongruency between context emotion and character emotion, we calculated the percentage of time out of the total amount of time in videos when the most selected emotion in the context-only ratings was different than that in the character-only ratings, which is on average 52.7% (SD: 20.6%) of the time in the videos. We then quantified the unique contribution of context versus character within this subset of video durations. We calculated the proportion of time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. The proportion of time when the context-only condition uniquely matched the fully informed condition was comparable to that of the character-only condition. This analysis suggests that the context effects reported in our study do apply to situations when the context is incongruent with the character. Error bars represent bootstrapped 95% CI. ** $p < 0.001$.

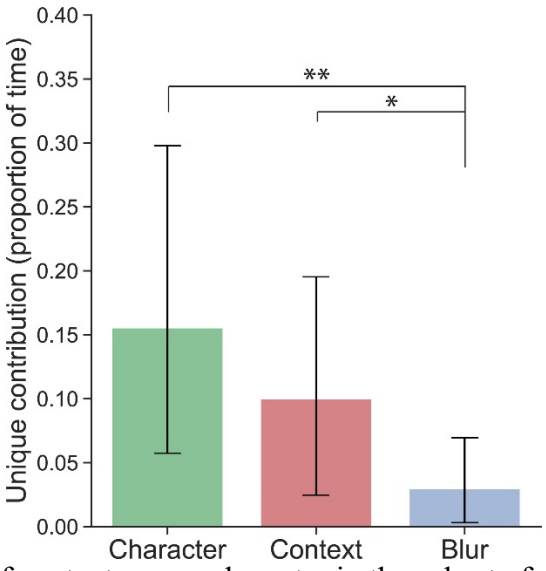


Fig. C3. Unique contribution of context versus character in the subset of video durations when the context emotion is in stark contrast with the character emotion. To quantify how often the emotion in the context-only condition is in stark contrast with the target (character-only) emotion, we calculated the percentage of time of the total amount of time in videos when the emotion in the context-only condition was of the opposite valence with the target emotion, which is on average 35.6% (SD = 31.4%) of the time in videos. We then quantified the unique contribution of context versus character within this subset of video durations. We calculated the proportion of time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. We found that the context (see red bar) still has a significant and unique contribution ($p < 0.01$, permutation test). This shows that the context effects that we found are not solely driven by video stimuli with similar context emotion and target emotion. Error bars represent bootstrapped 95% CI. * $p < 0.01$. ** $p < 0.001$.

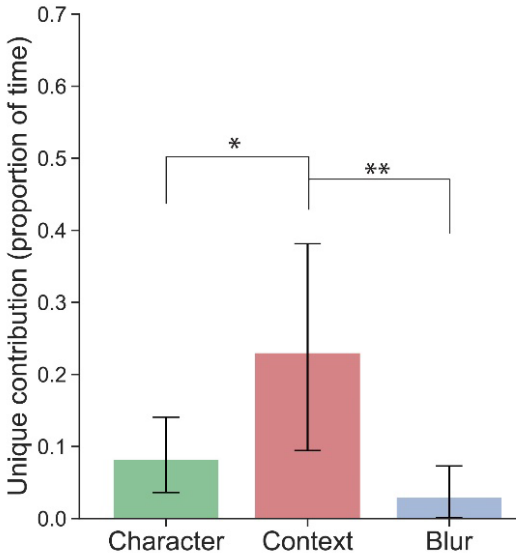


Fig. C4. Unique contribution of context versus character in the subset of 13 videos that are non-Hollywood movie clips (home videos or documentaries). The y-axis shows the proportion of

time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. The context-only condition was uniquely accurate 23% of the time (bootstrapped 95% CI: 9.3% - 38%; red bar). Surprisingly, we found that this is significantly more than the proportion of time when the character-only condition was uniquely accurate (mean: 8.2%; bootstrapped CI: 3.6% - 14.1%; green bar; $p < 0.01$, permutation test). These results suggest that the context may even be more influential in real-life situations when facial expressions are less exaggerated and diagnostic. Error bars represent bootstrapped 95% CI. * $p < 0.01$. ** $p < 0.001$.

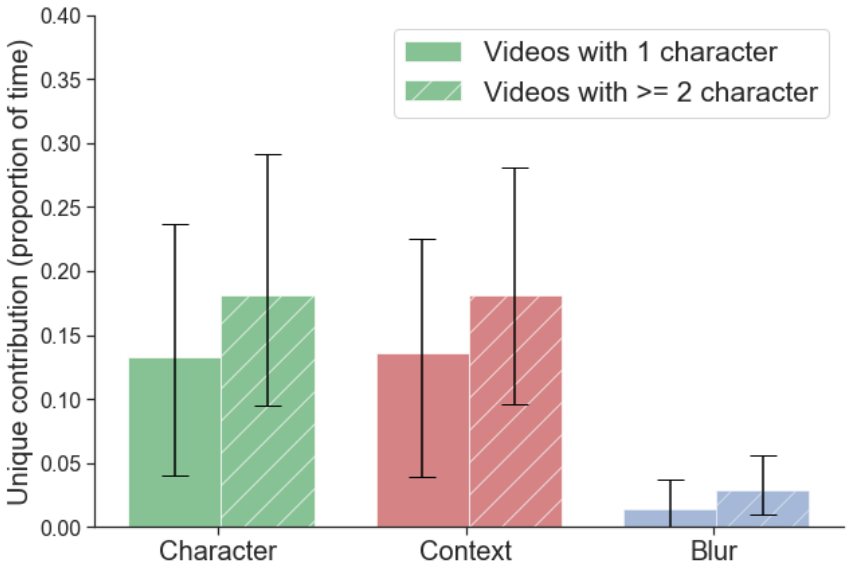


Fig. C5. Unique contribution of context versus character in the subsets of videos with or without other social agents present in the scene. In the stimuli we used, 9 of the videos (solid bars) show a single target character with no other social agent and the rest of the 24 videos (bars with stripes) have other social agents present. The y-axis shows the proportion of time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. Regardless of whether there were other visible social agents present in the videos or not, the unique contribution of the character is comparable to that of the context. We did not find any significant difference in our results that is attributable to the presence of social agents. Error bars represent bootstrapped 95% CI.

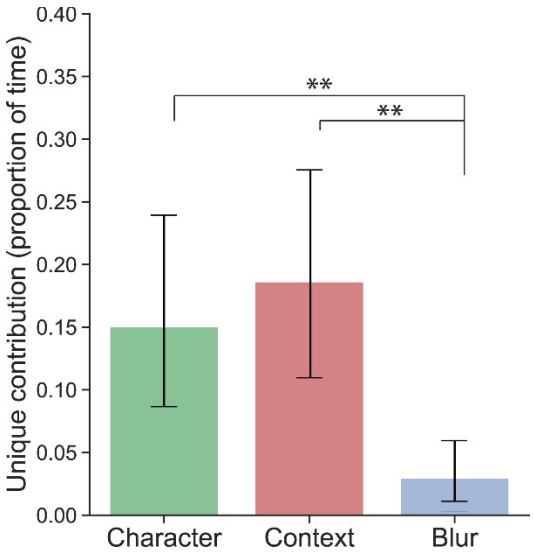


Fig. C6. Unique contribution of context versus character in the subset of trials participants rated the video clip to be not at all familiar. Participants rated their familiarity towards our video clips on a scale from 1 (Not at all familiar) to 5 (Extremely familiar) after viewing each video. In most trials (67%), participants rated the video clips to be not at all familiar (rating=1). Within this subset of data, we calculated the proportion of time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. We found that the context (see red bar) still has a significant and unique contribution ($p < 0.001$, permutation test). This shows that the context effects that we found are not solely driven by familiarity towards the video or the context. Error bars represent bootstrapped 95% CI. ** $p < 0.001$.

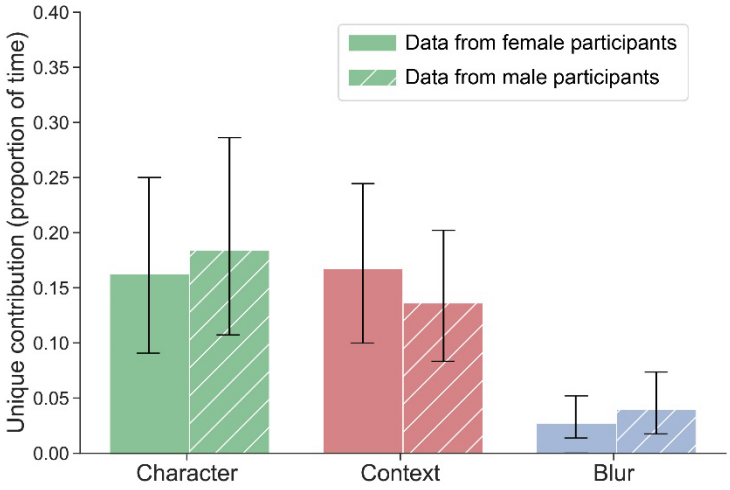


Fig. C7. Unique contribution of context versus character in the subsets of data comprised of female participants or male participants. On average, there are 30 female participants and 12 male participants rating each video. The y-axis shows the proportion of time out of the total amount of time in videos when the most selected category in the fully informed condition matches the most selected emotion in each condition but not any other condition. Regardless of participants' gender, the unique contribution of the character is comparable to that of the context.

We did not find any significant difference in our results that is attributable to the participants' gender. Error bars represent bootstrapped 95% CI.

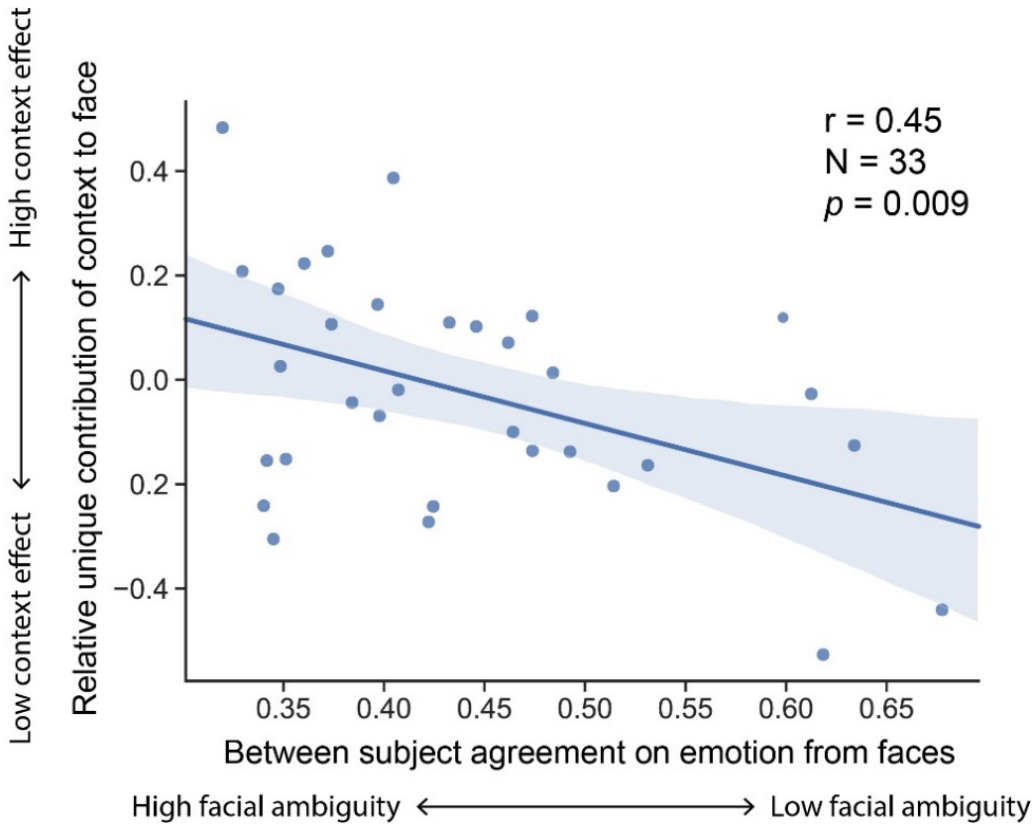


Fig. C8. Correlation between facial ambiguity and contextual contribution. We quantified the degree of ambiguity in facial expressions (on the x axis) using between subject agreement in the categorical character-only ratings. For every video, we calculated the mean proportion of participants choosing the most selected emotion category. If the facial ambiguity is low, more participants tend to agree on the same emotion category, and this proportion will be higher. To quantify the respective contribution of context versus face and body, we used another set of data from a previous published study collected on independent subjects and different rating scales (valence and arousal; Chen & Whitney, 2019). We used linear regression models to quantify the proportion of unique variance explained only by the character-only ratings or the context-only ratings for every video. Context effect (on the y axis) is then defined by subtracting the unique variance of character from the unique variance of context. A negative value on the y axis does not mean that context has no influence. It means that both character and context have influences but the influence of character is larger than that of context. We found a significant negative correlation (r = -0.45, p = 0.009) between measures of facial ambiguity and context effect across video clips. When facial expressions are more ambiguous, we rely more on contextual information for accurately recognizing emotion.