Emotion

# Inferential Emotion Tracking (IET) Reveals the Critical Role of Context in Emotion Recognition

Zhimin Chen and David Whitney

Online First Publication, December 28, 2020. http://dx.doi.org/10.1037/emo0000934

AMERICAN PSYCHOLOGICAL ASSOCIATION

Emotion


ISSN: 1528-3542

https://doi.org/10.1037/emo0000934

# Inferential Emotion Tracking (IET) Reveals the Critical Role of Context in Emotion Recognition

Zhimin Chen[1] and David Whitney[1, 2, 3]

[1] Department of Psychology, University of California, Berkeley

[2] Vision Science Program, University of California, Berkeley

[3] Helen Wills Neuroscience Institute, University of California, Berkeley

The ability to recognize others' emotions is critical for social interactions. It is widely assumed that recognizing facial expressions predominantly determines perceived categorical emotion, and contextual information only coarsely modulates or disambiguates interpreted faces. Using a novel method, inferential emotion tracking, we isolated and quantified the contribution of visual context versus face and body information in dynamic emotion recognition. Even when faces and bodies were blurred out in muted videos, observers inferred the emotion of invisible characters accurately and in high agreement based solely on visual context. Our results further 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. Visual context is an essential and indispensable element of emotion recognition: Without context, observers can misperceive a person's emotion over time.

*Keywords:* emotion, context, scene, facial expressions, emotion categories

*Supplemental materials:* https://doi.org/10.1037/emo0000934.supp



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 are 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 tech-

Zhimin Chen https://orcid.org/0000-0002-8075-072X

Library of videos and emotion ratings have been made available at https://osf.io/46rtw/.

Correspondence concerning this article should be addressed to Zhimin Chen, Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, United States. Email: chenzhimin@berkeley.edu

nique 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; Kret & de Gelder, 2010; Meeren et al., 2005; 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). Also, 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 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 men, age range 18–45, $M = 21$, $SD = 3.4$). Our participants comprised 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 the 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 .01.

### Stimuli

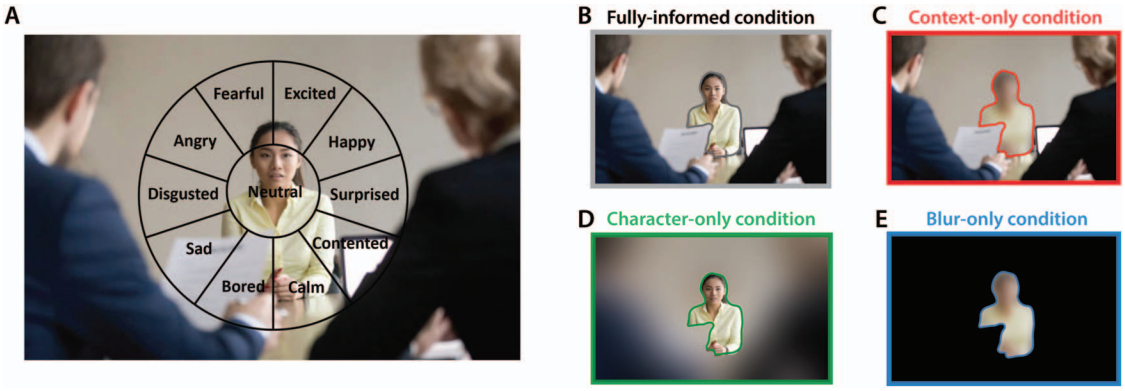
The same set of video stimuli was used in a previous study to collect ratings of valence and arousal (Chen & Whitney, 2019). The videos were gathered from an online video-sharing website based on the following criteria: (a) showing live action but not animation or monologue, and (b) the emotions/affect 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 came from home videos or documentaries. Eleven of the 12 home videos and documentaries included interpersonal contexts. The lengths of the videos ranged from 36 to 160 s.

To record real-time emotion judgments, we designed an emotion rating circle and superimposed it on top of the video (see Figure 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 six basic emotions are heavily skewed toward negative emotions with relatively high normative arousal, we added four other commonly studied emotion categories in order to achieve a relatively even distribution over normative valence and arousal. The nonbasic emotion categories include boredom (Rozin & Cohen, 2003), calmness (Cowen & Keltner, 2017), contentment (Keltner & Lerner, 2010), and excitement (Fredrickson, 1998; Shiota et al., 2017). 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): (a) "Neutral" was always at the middle of the circle. (b) 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. (c) 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 were nominally defined as the fully informed condition (see Figure 1B). To create stimuli in the context-only condition (see Figure 1C), we used state-of-the-art object segmentation algorithms (Mask R-CNN; He et al., 2020) and video-editing software (Adobe Premiere Pro CC) to selectively mask out a chosen target character in the video so

**Figure 1**
*Experimental Paradigm*




*Note.* (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 as shown in the online figure) when everything was visible. (C) In the context-only condition, participants tracked the blurred target (outlined in red as shown in the online figure) while the context remained visible. (D) In the character-only condition, participants tracked the visible target (outlined in green as shown in the online figure) while the context was blurred. (E) In the blur-only condition, participants tracked the blurred target (outlined in blue as shown in the online figure) while the context was masked completely by black pixels. Photo: reuse license purchased from iStock by Getty Images. The figure reprinted with permission. See the online article for the color version of this figure.

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 (see Figure 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 (see Figure 1E). Our library of videos and emotion ratings have been made available at https://osf.io/46rtw/.

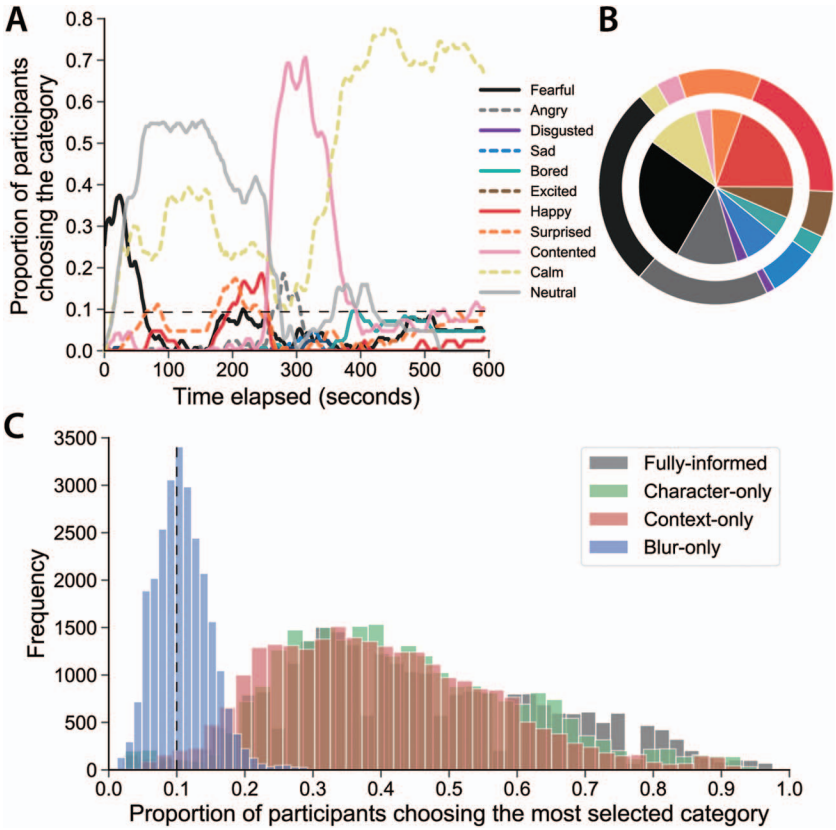
## Procedure

We used a similar procedure as Chen and Whitney (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: two thirds of the videos were from the blur-only condition and one third from 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 s 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 nonresponsive by calculating the longest duration that the participant had kept the mouse pointer in any single location. If the duration was longer than 20 s, 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 Figure 2A). To quantify consensus or between-subjects 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 ($r = .907$, $p < .001$; Bradley & Lang, 1999), which supports the representativeness of our video stimuli (see Figure 2B and online supplemental Figure S1 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 four different discrete emotions, and the dominant emotion was present for about 55.5% ($SD = 20.0\%$) of the video duration.

**Figure 2**
*Results: Between-Subjects Agreement*




*Note.* (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 ($p < .001$). (C) Between-subjects agreement in categorical emotion ratings across all videos. The percentage of people who chose the most selected emotion category in each 100 ms of the videos (not including "neutral," the default mouse position) is shown. Chance level (1/10 = 10%) is indicated by the dashed vertical line. See the online article for the color version of this figure.

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 nonneutral 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 ($M = 10.6\%$; $SD = 4.18\%$; Figure 2B, blue distribution). The between-subjects agreement in other conditions was substantially different from the blur-only condition, indicating that participants agreed more frequently than chance (Kolmogorov–Smirnov (K-S) test, $p < .001$, 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 (Figure 2B, red distribution), which is comparable to the fully informed condition (98.1%; Figure 2B, gray distribution) and the character-only condition (97.4%; Figure 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-subjects agreement. Visual context can be sufficient for inferring emotions that can be shared among perceivers.

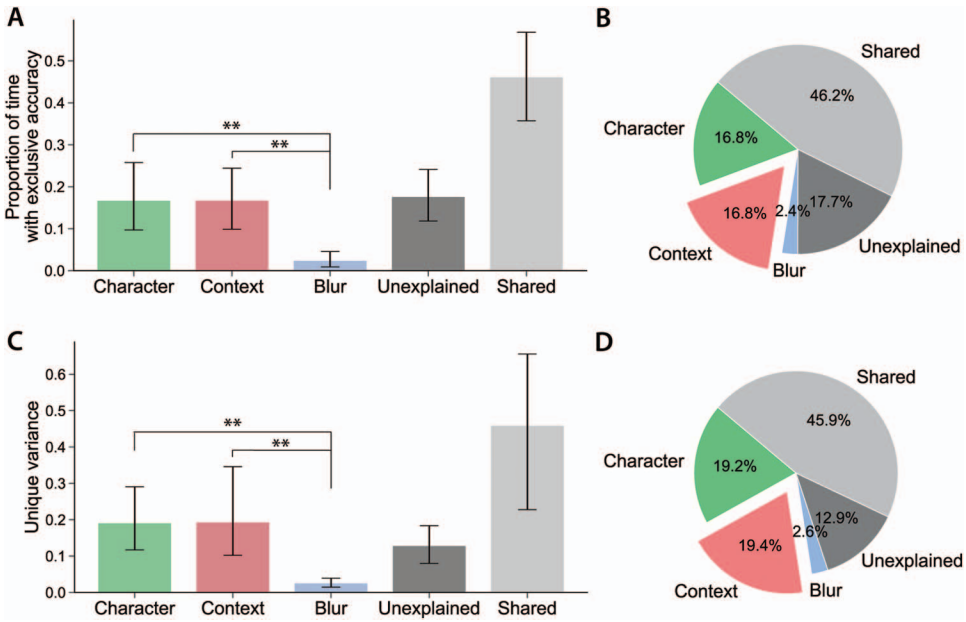
Is the context necessary to perceive and track emotion accurately, even when face and body information are already available? To answer this question, we compared the most selected emotion category for all time points across different conditions. We aggregated ratings across all participants and preserved 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 of 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 Figure 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 Figure 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 Figure 3A; $p < .001$, 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 ($p = .494$, permutation test) and was significantly larger than that explained only by the blur-only condition ($p < .001$, 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 the context-only, the character-only, or the blur-only variables 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%]; Figure 3C, red bar). This was significantly more than the unique

**Figure 3**

*Results: Unique Contribution of Context Versus Character*




*Note.* (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 as shown in the online figure), context-only emotion ratings (in red as shown in the online figure), and blur-only affect ratings (in blue as shown in the online figure). 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 1. Error bars represent bootstrapped 95% CI. ** $p < .001$. See the online article for the color version of this figure.

variance explained by the blur-only variables (mean: 2.60%; bootstrapped 95% CI [1.44%, 3.95%]; Figure 3C, blue bar; $p < .001$, permutation test). The proportion of unique variance explained by the character-only condition was 19.2% (bootstrapped 95% CI [11.7%, 30.0%]; Figure 3C, green bar), the magnitude of which was comparable to the unique variance explained by the context-only condition ($p = .510$, permutation test) and was significantly larger than that explained only by the blur-only condition ($p < .001$, 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 (see online supplemental Figure S3) and the subset with opposite valence between character and context (see online supplemental Figure S4). The contribution of context also remained significant in non-Hollywood movie clips (see online supplemental Figure S3), videos without any other social agent or character present (see online supplemental Figure S6), and videos that participants reported to be not at all familiar with (see online supplemental Figure S7). Similarly, we did not find a significant difference in context effects attributable to participants' gender despite the imbalance in sample sizes of gender groups (see online supplemental Figure S8). 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.

## 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 and 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 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-subjects 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 expressionless (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., 2012, 2017). 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 ($p < .05/11$, 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 (see Figure 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 is typically stationary and seems 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 time scales, 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 fea-

tures, 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 nonface 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 studies 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-subjects 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 (see online supplemental Figure S2).

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, 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 (AI) systems in commercial and educational settings that analyze emotions based solely on facial expressions (e.g., Affectiva.com; Microsoft Azure; Zeng et al., 2020). 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.

## References


Adelson, E. H. (2000). 24 Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 339–351). MIT Press.

Albright, T. D., & Stoner, G. R. (2002). Contextual influences on visual processing. *Annual Review of Neuroscience*, *25*, 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., 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

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

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*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

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

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

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

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

Clore, G. L., & Ortony, A. (2013). Psychological construction in the OCC model of emotion. *Emotion Review*, *5*(4), 335–343. https://doi.org/10.1177/1754073913489751

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 of the United States of America*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

de Gelder, B. D., & 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). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199559053.013.0027

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

Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology*, *2*(3), 300–319. https://doi.org/10.1037/1089-2680.2.3.300

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)E7900–569. https://doi.org/10.1037/h0057585

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

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, *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

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). Oxford University Press.

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

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

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

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

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). Guilford. Retrieved from https://psycnet.apa.org/fulltext/2008-07784-013.pdf

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

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

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

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

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. https://doi.org/10.1037/emo0000445

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. https://doi.org/10.1017/CBO9781316275672.008

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. *American Psychologist*, *72*(7), 617–643. https://doi.org/10.1037/a0040456

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*(36), 12412–12424. https://doi.org/10.1523/JNEUROSCI.4822-14.2015

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

Wieser, M. J., & Brosch, T. (2012). Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*. Advance online publication. 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.

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*. Advance online publication. https://doi.org/10.1109/TVCG.2019.2963659

Received January 30, 2020
Revision received August 31, 2020
Accepted September 14, 2020 ■