Context in Emotion Recognition: A Replication and Extension Study

Alexandra Abillar, Steven-Jethro Chan, Jason Huang, Christiana Kang, Chania Kim, Sai Keerthana Puvvula, Amber Wei

Cognitive Science and Psychology Undergraduate Laboratory @ Berkeley

Undergraduate Student Mentors: Megan Cervera, Yifei Chen

Graduate Student Mentor: Jefferson Ortega

University of California, Berkeley

CONTEXT IN EMOTION RECOGNITION: A REPLICATION AND EXTENSION STUDY

Abstract

Emotional recognition plays a crucial role in everyday interpersonal communication, mental health, and technology application. Recently, it has been proposed that context is not only sufficient but also necessary for emotional inference and thus contributed to the understanding of the process of emotion recognition (2019). This paper focused on replicating and extending a previous study that investigated emotional inference of characters in visual-only videos in the dimensions of valence and arousal under either fully informed or context-only conditions (Chen & Whitney, 2019). The replication study yielded similar results to experiment one: participants inferred the emotion of the invisible characters with high accuracy and in high agreement based on context. However, the target and partner characters' arousal ratings across the time differed. As an extension to the accuracy of emotion inference, how individual inference concurs with the average inference was analyzed. Participants' inferences were more accurate (concurred more with the average inference) for arousal than valence. It has shown that emotional recognition is more accurate with stronger emotion.

Introduction

Emotion recognition — understanding one's own and other's emotions — is vital in an individual's social interactions (Olsson & Ochsner, 2008). For example, one may recognize emotion through subtle facial cues, changes in body language, and contextual evidence with the involvement of motor neurons (GALLESE et al., 2004), allowing society to interact in mutual understanding and benefit. More recently, the ability to recognize and interpret emotions played a key part in defining emotional intelligence, an essential part in diagnosing various cognitive illnesses such as schizophrenia (Kohler et al. 2010) or autism (Harms et. al 2010). Tests that diagnose (Hooker & Park 2002; Penn et al. 2000, Addington & Addington 1998) cognitive disorders like schizophrenia or autism implement images of facial expressions for participants to identify.

Emotion recognition through facial expressions

Emotion recognition through decontextualized facial expression — identifying emotion through facial cues involves noticing the subtle differences in the movement of the eyes, lips, cheeks, eyebrows, and nose — is thought to be one of the primary methods of emotion recognition. Tests like the previously mentioned diagnostic procedures can include showing participants various decontextualized faces with a person smiling, frowning, confused, etc (Hooker & Park, 2002), given in the form of an exam that is scored on accuracy. However, as mentioned in Olsson & Ochsner 2008, nonverbal communication is difficult to identify, especially in more complicated situations. It's clear that emotion recognition is performed rarely through isolated facial expressions only and thus includes body context (Aviezer et al., 2011).

Emotion recognition through body language

As mentioned, emotion recognition is often associated with body language and posture — the movement of arms, legs, head, etc, can all contribute to what is understood as someone's emotional state. For instance, a tilted head may indicate surprise or confusion while stomping may be an expression of anger. Again, while more challenging to manipulate, body language can still be difficult to interpret without contextual clues. The existence of mirror neurons (GALLESE et al., 2004) may help to identify decontextualized body language by mirroring the actions of the subject. However there is still more to the picture than just facial or body expression, but environmental context as well (Aviezer et al., 2008).

Context in emotion recognition

The original paper by Chen and Whitney works to connect facial cues, body language, and contextual evidence as a unified method of recognizing emotion in others. Contextual clues may include the setting, the visual scene, facial expressions in other characters, and other verbal and nonverbal cues from the background. As Chen and Whitney explain, seemingly simple facial expressions out of context may be misinterpreted incorrectly: a smile can hide nervousness and anxiety, something that may be impossible to detect without more information.

Chen and Whitney sustains that context is just as important as facial and body language by testing and comparing valence and arousal ratings in videos with a subject's face hidden in various scenarios. Valence and arousal can be defined by the negativity/positivity and the intensity of an emotion, respectively. For example, excitement has positive valence and high arousal while indifference may be measured by neutral valence and low arousal.

It explores the relationship between visual context and emotion perception while acknowledging the multiple variables contributing to emotion rating (Chen & Whitney, 2019). To understand how visual context affects emotion perception, Chen and Whitney utilized affective tracking, a method that tests how well a person can infer emotion from contextual cues. They focus on valence and arousal perceptions due to the fact that it captures the most variance in emotion rating. Overall, they found that participants can discern emotion from unique information only found in the background context. Without the addition of a face, the visual context appears to be enough. This correlation aids in the notion that emotion recognition includes multiple variables, allowing for further research exploring how daily life situations affect one's emotion recognition.

Extension of Chen & Whitney; Effect of emotional intensity on recognition accuracy

This present study aimed to replicate and extend the findings of Chen and Whitney to investigate the effect of context in emotion recognition. By replicating the conditions set by Chen and Whitney, we confirmed that dynamic visual context may play a role in emotion recognition, demonstrating that it does not rely solely on information from facial expressions and body postures. After reproducing the conditions Chen and Whitney created for their participants, these findings allow us to better examine the finer details of background context and emotion recognition.

This study will expand on the correlation between visual cues and emotional perception. Since viewers successfully determined a person's emotions in scenarios within a video, it is important to observe how this may affect emotion recognition in daily life. Therefore, this study allows us to explore our hypothesis: both aspects of emotion perception (arousal and valence) are

used in contextualized environmental conditions between the context group and the baseline group.

Methods

Data and Participants

The replication and extension studies used data from experiment one in the original study. In experiment one, 33 participants were asked to infer and track the emotion of the invisible target character and visible partner character. Given the interest in how emotion detection may vary under different circumstances, fully informed, context only, and fully informed-context combined groups were included. Specifically, the data is composed of arousal and valence ratings of each participant per timestamp continuously across the experiment.

Measures and Procedure

Participants measured their inferred emotions via a two-dimensional (2-D) valence-arousal grid with valence and arousal as the axes. Valence can be described as the excitement level, ranging from low to high, while arousal is described as a pleasantness level, defined along a continuous line between negative and positive. The 2-D grid separates emotion into independent dimensions that are cited to be both independent and predictive of one another [Bestelmeyer et. al 2017]. Participants put a mouse pointer at the location within the grid, which they believe best quantifies their emotions at that time stamp.

During the experiment, participants viewed a silent Hollywood movie clip and used the aforementioned 2-D valence-arousal grid to continuously report the affect of a chosen character in the video. Participants are randomly assigned to two conditions: inferred or fully informed. The inferred condition involves the target character occluded by a Gaussian blurred mask while

the partner was visible. The fully informed, or baseline, condition involves everything visible. Participants are asked to infer and track the affect of the target character either in inferred condition or fully informed condition. Only participants assigned to the fully informed condition infer the partner character. In this way, there were six datasets.

Replication Data Analysis

For the replication study, analysis was performed in Python. All ratings are first standardized using z-score. The first hypothesis tested if context is sufficient for emotion recognition; inferences of target and partner characters, in inferred and fully informed conditions, show a similar pattern. To address Hypothesis 1, standardized valence/arousal ratings are calculated as the average ratings per timestamp across all participants in each group. Standard errors were also calculated considering variances among individual participants.

To confirm whether there is a high intersubject agreement in the inferred affect ratings of the invisible character for valence and arousal or not (Hypothesis 2), average pairwise correlation coefficients were calculated. Baseline and context only ratings were combined for each video clip. Average pairwise correlation coefficient was calculated by the following steps. For each video clip, every possible two pairs were selected to calculate their correlation coefficient. The pairwise correlations were averaged in one video clip. Correlations of all video clips for each dimension of affect were further averaged to gain the average total correlation of valence and arousal. All averaged correlations were computed by applying Fisher Z-transformation on all individual correlations, averaging the transformed values, and then transforming the mean back to Pearson correlation.

The single-subject correlation correlations are normalized by their ceiling. Using data from the reliability test, for each video clip, correlations were calculated between the first test

and retested to find the correlation between individuals and themselves for valence and for arousal. These correlations can represent the ceiling values because individuals are most correlated with themselves. Abnormal correlations that are equal to and larger than 0.99 were removed. For each dimension of effect, average correlation of each subject across all video clips was calculated and later averaged the correlation across all subjects. In this way, we gained the ceiling of arousal and the ceiling of valence.

The third hypothesis is that inference under context only condition would have comparable accuracy with inference under fully informed condition. To quantify IAT accuracy, Pearson correlations between inferred affect ratings (context only) and fully informed rating (baseline) of the target are calculated for each group. Average correlation of each group was calculated separately. To figure out whether the background matters in participants' tracking process, partial correlation coefficients between inferred affect ratings (context only) and fully informed rating (baseline) by controlling for fully informed affect ratings of the partners were calculated. Using baseline, context- only, and context- only control data, partial correlations of valence and arousal were respectively calculated. All average partial correlations were applied with Fisher Z-transformation first.

Extension Data Analysis

For the extension study, analysis was performed in Python. Rating and accuracy were used in the following analyses. Rating refers to the emotional strength expressed by a video clip, measured by the average rating magnitude of all participants across the duration of one video. Accuracy refers to the general agreement about the emotional strength of one video among all participants. The accuracy was calculated by the average correlation between participants' ratings and the overall average rating for each video and character combination for both valence

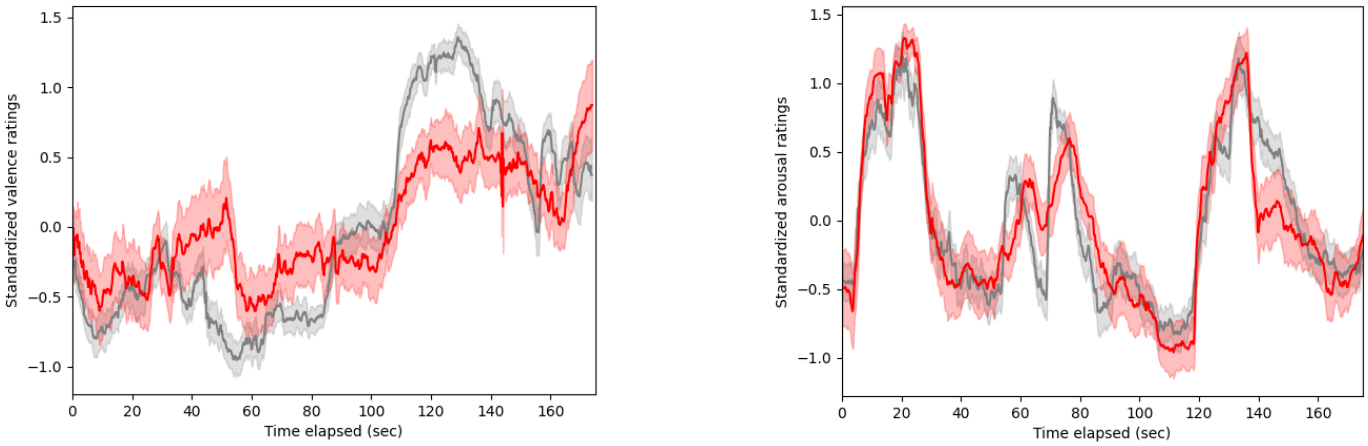
and arousal. All average correlations were computed by first applying Fisher Z-transformation on all individual correlations between each participant's rating and the rating of the video clip, averaging the transformed values, and then transforming the mean back to the original scale. Data was first analyzed as a whole regardless of which condition data originated, separated by the aspects of emotions (arousal vs. valence). Further comparison was made to examine whether there is a difference between the baseline and context groups. For arousal and valence, the rating and accuracy of the 26 video clips were calculated. Pearson correlation coefficient and spearman's rank correlation coefficient were computed to quantify the relationship between rating and accuracy.

Results

Replication Results

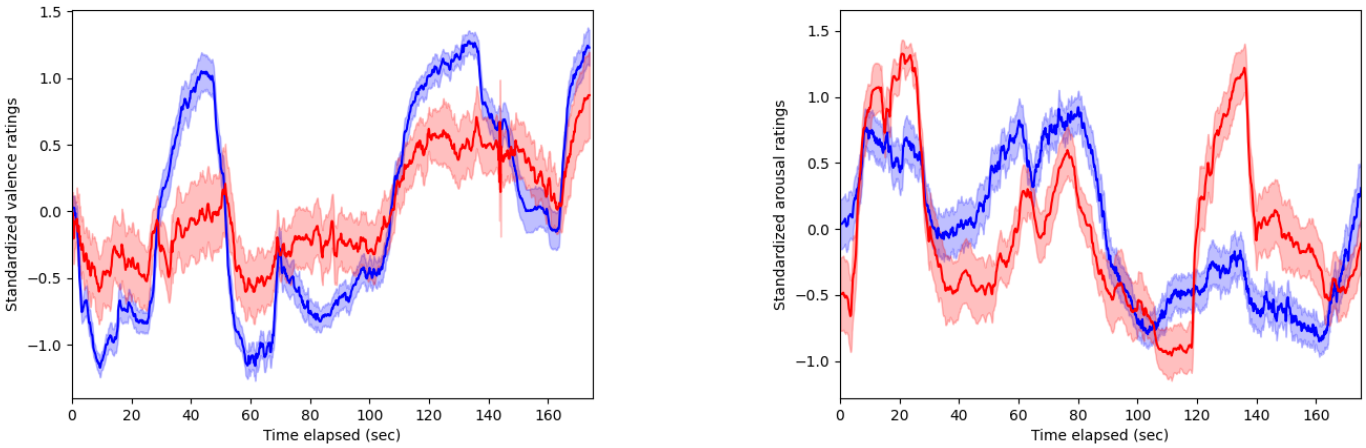
Figure 1

Standardized valence and arousal ratings for target and partner characters



Tracking Invisible Target Character in Inferred Condition vs.

Tracking Visible Target Character in Fully-informed Condition



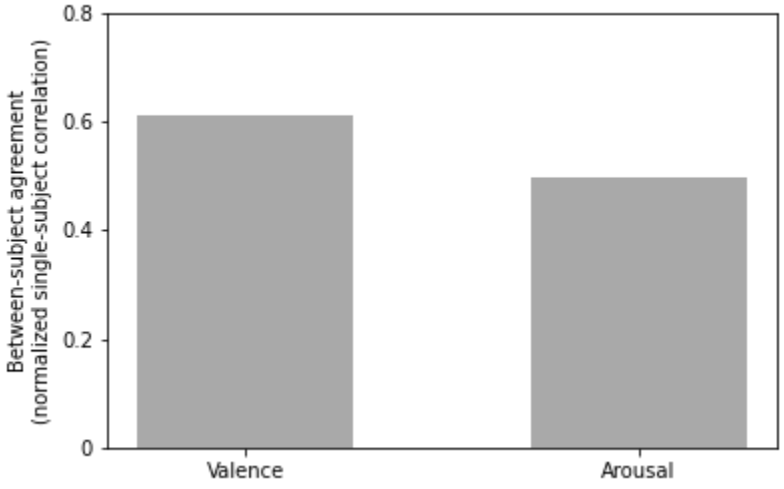
Tracking Invisible Target Character in Inferred Condition vs.

Tracking Visible Partner Character in Fully-informed Condition

Note. The light area represents the standard error obtained from the standard deviation of the sample over the square root of the sample size.

Figure 2A

Between-subject agreement (normalized single-subject correlation)

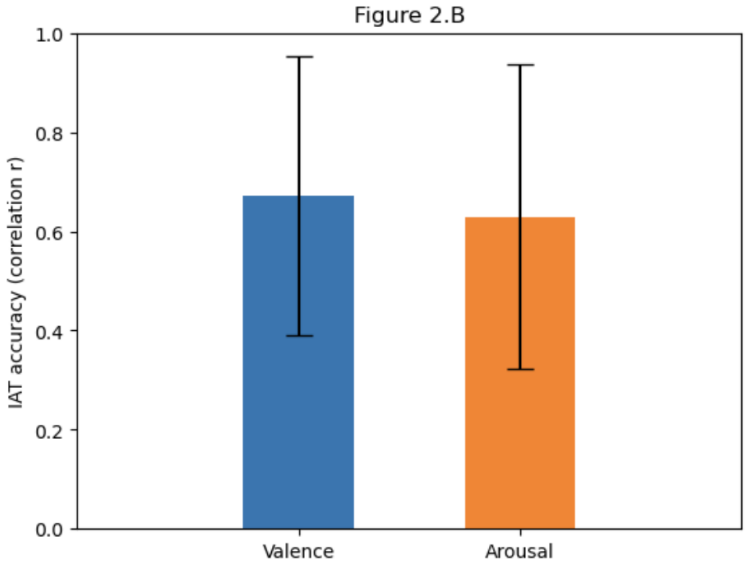


The level of between-subject agreement is relatively high shown by normalized single-subject correlation of 0.61 for valence and 0.49 for arousal.

Our calculations show a high similarity between inferred affect ratings (context only) and fully informed rating (baseline) of the target, with 0.63 and 0.67 Pearson correlations for arousal and valence, respectively (Figure 2b).

Figure 2B

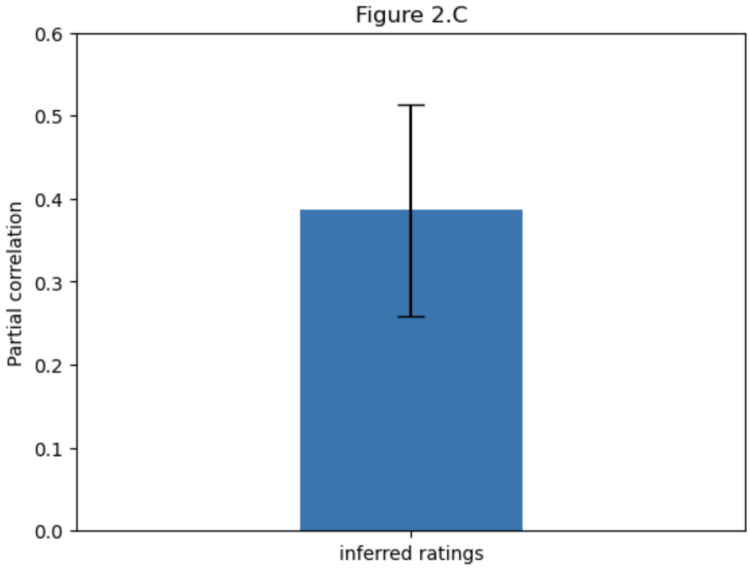
IAT accuracy



After that, we average these two arrays and finally get a value of 0.39 (Figure 2c). This suggests that when participants try to infer the invisible target, not only do visible partners help make an inference, but backgrounds also play an important role there.

Figure 2C

Partial Correlation

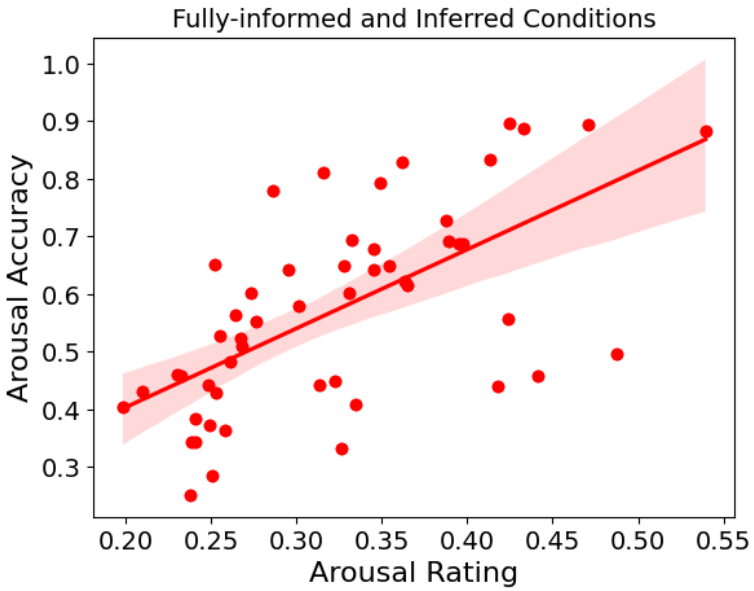


Extension Results

For arousal (when considering the fully-informed and inferred conditions together), there is a strong positive relationship between rating and accuracy within the fully-informed condition (Figure 3). Pearson correlation coefficient is $r(50)=0.63$, $p < 0.001$. Spearman's rank correlation coefficient is 0.66, $p < 0.001$.

Figure 3

Arousal Accuracy versus Rating with a linear regression model fit

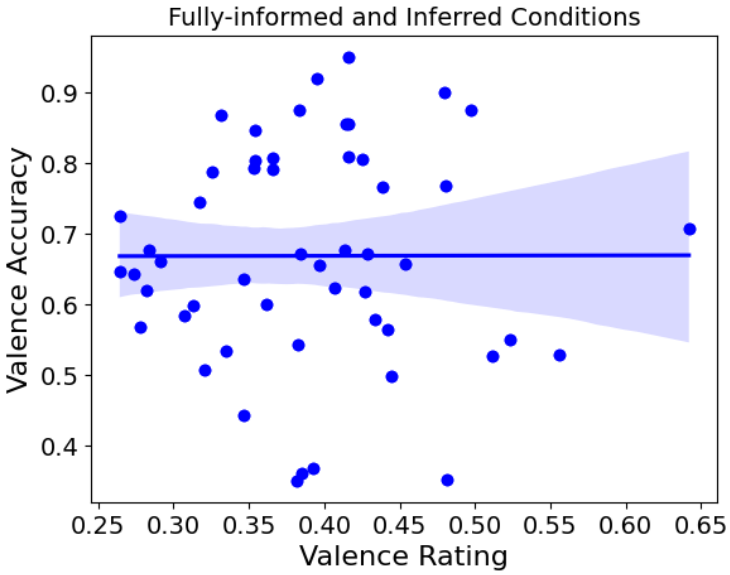


Arousal Accuracy versus Rating with a linear regression model fit

When considering the fully-informed and inferred conditions together, there was no statistically significant correlation between rating and accuracy for valence (Figure 4). Non-significant correlations were seen in both the Pearson correlation coefficient, $r(50) = 0.002$, $p > 0.05$, and Spearman's rank correlation coefficient, 0.02, $p > 0.05$.

Figure 4

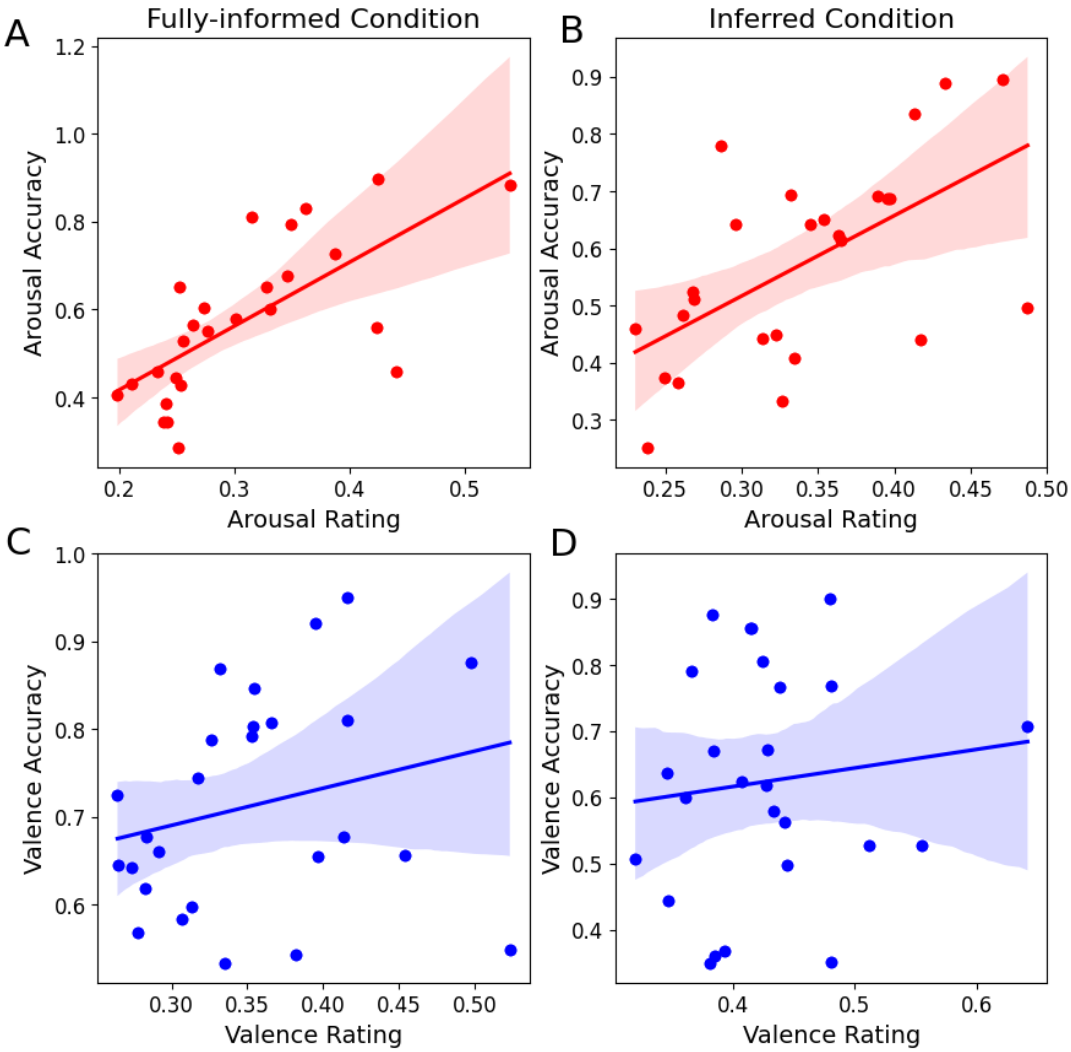
Valence Accuracy versus Rating with a linear regression model fit



For arousal, a strong positive relationship between rating and accuracy exists within the fully-informed condition (Fig 5a). Pearson correlation coefficient is r(24)=0.69, $p < 0.001$. Spearman's rank correlation coefficient is 0.77, $p < 0.001$. For arousal, there is a moderate, positive relationship between rating and accuracy within the inferred conditions (Fig 5b). Pearson correlation coefficient is r(24)=.59, p<.01. Spearman's rank correlation coefficient is .56, p<.01. For valence, there is no significant correlation between rating and accuracy within the fully-informed condition (Fig 5c). Pearson correlation coefficient is r(24)=0.24, $p>0.05$. Spearman's rank correlation coefficient is 0.33, $p>0.05$. For valence, there is no significant correlation between rating and accuracy within the inferred conditions (Fig 5d). Pearson correlation coefficient is r(24)=.11, p>.05. Spearman's rank correlation coefficient is .08, p>.05.

Figure 5

Arousal and Valence Accuracy versus Rating within each condition



Discussion

Our replication of the study confirmed that participants use information about the background context, independent of any face information, to accurately register and track effect. Even when no face information is present, the visual context is sufficient to infer valence and arousal over time. Different observers agree about the affective information provided by the context. This confirms our hypothesis that the absence of facial information might shape and

influence the actual perception of emotional signals. Our results provide clear evidence that the context by itself is both sufficient and necessary for accurate judgments of the perceived effect 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 effect of faces and bodies. Observers actively derive information from contextual information and face and body information and combine them in a context-specific way in real time. Importantly, these substantial contextual influences were observed with a range of different video stimuli, including those with and without interpersonal interactions, with posed or spontaneous facial expressions, and with staged or natural scenes.

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

actively deriving information from context because facial expressions in real life are often absent, ambiguous, or nondiagnostic of emotion (40, 41). In summary, we can better understand the perceptual and neural mechanisms of emotion perception if we incorporate and measure the critical role of contextual information. Our technique allows for this.

The results indicated a significant positive correlation between rating and accuracy in predicting arousal. This suggests that as the rating increases, so does the accuracy of the arousal prediction. In other words, the methods employed in the study were more effective in identifying and quantifying the intensity of emotions (arousal) experienced by the unseen individuals. However, the results did not show a significant correlation between rating and accuracy in predicting valence, which refers to the positive or negative nature of the emotions. This implies that the methods used in the study were less effective in determining whether the emotions experienced by unseen individuals were positive or negative. In summary, the study highlights that the employed techniques were more successful in predicting the intensity of emotions than in identifying their positive or negative nature.

We suspect that preference is a major influencing factor resulting in the deviation between the correlation of our arousal and valence data. The result of our experiment suggests that there is a greater accuracy in reaction to intensity but significantly less for positive or negative emotion. This is parallel to the nature of distinction (in this case taste) between individuals. The difference will then invoke bias particularly through 2 aspects: genre bias and scene composition bias. Different movie genres can be associated with different levels of valence and arousal, but certain genres are difficult to be directly associated with arousal or valence. For example, one of the movie clips is about “a little boy who discovers he has been left home alone, and he takes advantage of his freedom and plays in the house”. While one may perceive it as

positive by relating to the child, a counter feeling can also be applied if the audience is a parent. However, whether we take our stance as the child or some parent, it is clear that the clip invokes a certain degree of emotions. The composition of a scene can also bias the viewer's emotional response. The use of lighting, camera angles, and other visual cues can influence the level of arousal and valence a scene elicits.

Limitations of this extension study include genre bias, music bias, and scene composition bias. Emotion is a complicated concept and the valence and arousal ratings are not the most accurate measures of a level of emotion. Genres of the scene can affect the emotion produced, such as horror movies creating suspense and fear and romantic movies creating feelings of joy and happiness. It is expected that different genres of scenes may produce different correlation levels between emotional intensity and ratings between participants, even if the valence/arousal ratings of each scene are similar. The same is true for music and scene composition. Both heavily affect the *type* of emotion that is produced, which may produce varying correlations between intensity and accuracy. A potential extension study can test these very concepts by comparing between different genres, music types, and scene compositions. Additionally, this study did not account for potential confounding variables, which include age, neurological disorders, and psychiatric disorders. Previous studies have shown that people with these conditions tend to perceive emotions differently. In a study, these individuals with disorders specific symptoms showed abnormal brain activity such as abnormal cognition and abnormal emotion processing (Lee et al., 2016).

Regardless of the conditions, there is a positive correlation between the accuracy and rating in arousal whereas valence exhibits little correlation. Our findings only support part of Chen and Whitney’s conclusion as in the inferred condition, we observed that visual contact is

sufficient to infer arousal only. However, the results still depict a dependence on environmental context for recognition (Aviezer et al., 2008) since the correlation for arousal and valence were constant in the fully-informed and inferred conditions. Overall, our results demonstrate that emotion recognition through environmental context is partially dependent on arousal whereas the accuracy and rating for valence must come from another source of information.

Acknowledgements

References

Addington, J., & Addington, D. (1998). Facial affect recognition and information processing in schizophrenia and bipolar disorder. Schizophrenia Research, 3, 171–181. https://doi.org/10.1016/s0920-9964(98)00042-5

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

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, Disgusted, or Afraid? Psychological Science, 7, 724–732. https://doi.org/10.1111/j.1467-9280.2008.02148.x

GALLESE, V., KEYSERS, C., & RIZZOLATTI, G. (2004). A unifying view of the basis of social cognition. Trends in Cognitive Sciences, 9, 396–403. https://doi.org/10.1016/j.tics.2004.07.002

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, 3, 290–322. https://doi.org/10.1007/s11065-010-9138-6

Hooker, C., & Park, S. (2002). Emotion processing and its relationship to social functioning in schizophrenia patients. Psychiatry Research, 1, 41–50. https://doi.org/10.1016/s0165-1781(02)00177-4

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2009). Facial Emotion Perception in Schizophrenia: A Meta-analytic Review. Schizophrenia Bulletin, 5, 1009–1019. https://doi.org/10.1093/schbul/sbn192

Lee, S. A., Kim, C.-Y., & Lee, S.-H. (2016). Non-Conscious Perception of Emotions in Psychiatric Disorders: The Unsolved Puzzle of Psychopathology. Psychiatry Investigation, 2, 165. https://doi.org/10.4306/pi.2016.13.2.165

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

Penn, D. L., Combs, D. R., Ritchie, M., Francis, J., Cassisi, J., Morris, S., & Townsend, M. (2000). Emotion recognition in schizophrenia: Further investigation of generalized versus specific deficit models. Journal of Abnormal Psychology, 3, 512–516. https://doi.org/10.1037/0021-843x.109.3.512