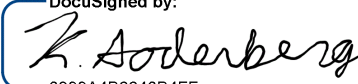


Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

DocuSigned by:
Signature: 
0399A4B6640B4FF...

Katherine Soderberg

Name

5/16/2023 | 1:10 PM EDT

Date

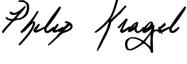
Title The Role of Experience in Emotion Understanding

Author Katherine Soderberg

Degree Master of Arts

Program Psychology

Approved by the Committee

DocuSigned by:

E39E7EB16272405...

Philip Kragel

Advisor

DocuSigned by:

685971CA57FC4EC...

Patricia Brennan

Committee Member

DocuSigned by:

E5E6836C7DCE4A1...

Michael Treadway

Committee Member

Accepted by the Laney Graduate School

Kimberly Jacob Arriola, PhD, MPH

Dean, James T. Laney School of Graduate Studies

The Role of Experience in Emotion Understanding

Katherine M. Soderberg

Department of Psychology, Emory University

Abstract

Inferring others' emotions from their varied expressions is a powerful human capacity that is particularly important in close relationships. Despite the fact that accurately interpreting the emotions of familiar others enhances well-being, relatively little research has explored the effect of prior experience on emotion understanding. It is possible that a general mental model connecting expressions to emotions supports emotion inference; alternatively, perceivers might develop person-specific mental models that map idiosyncratic expressions onto emotions based on past experience with individual targets. To arbitrate between these two possibilities, we manipulated prior exposure by showing subjects clips (which varied in modality) of characters from the show *Friends*. We also capitalized on subjects' past experience with those characters, categorizing those who had seen the show as "experts" and those with minimal exposure as "naïve." Then, we asked subjects to infer the emotions characters were feeling in a set of short audiovisual clips. Using Gaussian mixture modeling, we clustered subjects' emotion ratings into a "ground truth" cluster solution. We classified held-out expert and naïve subjects' ratings according to this ground truth solution using a support vector machine classifier with a radial basis function kernel to examine if there were differences between expert and naïve groups. Interestingly, classification generalized across expert and naïve groups, suggesting that a general model of emotional expression may be sufficient to explain emotion inference in some contexts.

THE ROLE OF EXPERIENCE IN EMOTION UNDERSTANDING

BY

KATHERINE M. SODERBERG

THESIS

Submitted to the Department of Psychology

Laney Graduate School, Emory University

In Fulfillment of the Requirements

For the Degree of Master of Arts

June 2023

Contents

| | |
|-----------------------------|----|
| Introduction..... | 5 |
| Methods..... | 10 |
| Results..... | 15 |
| Discussion..... | 18 |
| References..... | 24 |
| Supplemental Materials..... | 29 |

The Role of Experience in Emotion Understanding

Humans expertly interpret others' emotional states from dynamic, multimodal signals. Presented with an array of complex, context-dependent cues, the human brain integrates emotion signals to arrive at an inference of how a target might be feeling. This ability is particularly crucial in the context of close relationships; correctly inferring the mental state of another person, which is known as empathic accuracy, has been consistently linked to relationship satisfaction (Sened et al., 2017). Strong relationships provide social support, which promotes physical and psychological well-being (Holt-Lunstad et al., 2010). Given that understanding the emotions of close others is especially consequential, it is important to know how past experience with a target influences the way perceivers make emotion inferences. However, our understanding of the process of emotion recognition is based on studies that predominantly use unfamiliar targets; less is known about how emotion recognition is affected as perceivers learn the idiosyncrasies of targets' expressions. One promising idea posits that perceivers build person-specific mental models that are strengthened by experience with a target; alternatively, they might use a general mental model that maps sensory signals to emotion, regardless of who is making the expression.

When probing how the brain might arrive at an emotion inference from sensory input, it is helpful to consider the broader context of the predictive mind. In this theory, which is supported by evidence spanning neural and behavioral levels of analysis, the brain is constantly estimating priors, and comparing these predictions to sensory reality (Hohwy, 2013; Rao & Ballard, 1999). A predictive mind, therefore, should have a model of others' emotions and how they are expressed. Amongst the possibilities for how such a model could function, there are two extremes. One is a general mental model, in which sensory signals are linked to emotions

regardless of the person who is making the emotional display. The other is a collection of person-specific mental models, in which the idiosyncrasies of an individual's expression are learned and used to predict their emotions in the future. Evidence from other domains of social cognition suggests that people use person-specific mental models when thinking about others. In one study where subjects were asked to predict the likelihood that a target would transition from one emotional state to another, they were more accurate when predicting a friend's compared to a stranger's emotions and used person-specific knowledge to do so (Zhao et al., 2020). Person-specific mental models have also been proposed in the context of theory of mind: in a neuroimaging study in which people evaluated political figures, Welborn & Lieberman (2015) found increased dmPFC activation for subjectively better-known targets as well as targets judged to be idiosyncratic, compared to less-known targets. Although there are a range of possible explanations for this finding, the authors conclude that mPFC might tune mentalizing to individual-specific representations. This emerging evidence suggests that social cognitive processes may involve person-specific predictions based on past experiences, rather than solely relying on a general mental model. It is an open question whether a similar pattern is seen in the case of emotion recognition, and if so, how person-specific mental models are implemented in the brain.

To understand the emotions of familiar people, the brain must process two essential components: identity and expression. These components are commonly and reliably signified by faces, making face perception an ideal starting point for understanding them. A prominent account from the face-perception literature defines two neural systems for processing faces (Haxby et al., 2002). One, in the ventral stream, processes invariant features, which are important for recognizing the identity of an individual. The other, anchored in the superior temporal sulcus

(STS), processes dynamic features, which underlie the representation of emotional expression. In this account, these systems interact, yet are specialized and dissociable. However, alternative interpretations of the data challenge the idea of dissociable pathways and posit that expression and identity processing interact in early stages of these pathways, potentially in STS (Calder & Young, 2005; Calder, 2011). This interaction may support the recognition of emotion for familiar others. If person-specific mental models are in fact undergirding emotion understanding, inputs from the identity system might influence representations in the expression system, allowing for a more precise understanding of the familiar person's emotion.

Indeed, the central node of the expression system, the STS, has been shown to be modulated by familiarity outside the context of emotion. Increased activation in STS has been observed when people view personally familiar faces compared with faces of familiar targets with whom there is less experience (Gobbini et al., 2004; Leibenluft et al., 2004). Notably, patterns in STS trained on representations of faces were able to discriminate between voices of the same individuals, indicating that this region may contain multimodal representations of the identity of specific targets (Tsantani et al., 2019). This evidence that STS is sensitive to identity lays important groundwork for the possibility that identity alters the processing of emotional expressions, allowing for the representation of individual idiosyncrasies.

The emotion recognition literature has revealed a great deal about how the process of emotion inference unfolds for unfamiliar targets. Along with facial expressions, emotions are inferred from a wide range of signals, including body posture, vocal tone, semantic content, and context. These initial sensory signals are processed by corresponding sensory areas in visual and auditory cortex (Harry et al., 2013; Ethofer et al., 2009); facial and vocal expressions also lead to activity in somatosensory cortex (Kragel & LaBar, 2016). Evidence suggests that these

multimodal signals are integrated in STS (Campanella & Belin, 2007; Peelen et al., 2010; Watson et al., 2014). This integration likely allows for multimodal representations of emotion recognition, but it is unclear exactly how each modality contributes to the integrated emotion judgment. Several behavioral studies in which expressions are presented in different modalities have revealed that verbal information (from audio or text) is most important for accurately inferring a target's emotions (Gesn & Ickes, 1999; Hall & Schmid Mast, 2007; Jospe et al., 2020). Semantic content is undoubtedly important when making judgments about others' emotions; it is less clear how this information gets integrated with other emotion signals, and how this might be done when different modalities lead to competing predictions.

A smaller number of studies have examined how emotion understanding is affected by experience. It is well-established that prior experience with a target improves the accuracy of emotion judgments (Stinson & Ickes, 1992; Thomas & Fletcher, 2003). How this is accomplished is more uncertain. One explanation is that person-specific mental models allow for more precise inferences about what a target is feeling. This could be accomplished if information about a target's identity influenced emotional expression processing, as possibly occurs in STS. One neuroimaging study found separate fMRI adaptation effects for identity and expression in distinct subregions of STS (Winston et al., 2004). It is possible that in the case of recognizing the expressions of familiar others, the expression and identity tracking taking place in STS allow for person-specific emotion signals to be learned. It is also possible that cross-talk between identity and expression takes place in later regions of the ventral visual stream, including the fusiform face area. In another study, researchers recorded intracranially from STS and ventral visual regions as subjects undergoing epilepsy monitoring viewed emotional faces and found that they

were better able to decode expressions from the activity of ventral visual regions compared to STS (Tsuchiya et al., 2008).

One way to investigate how identity and emotion recognition interact in the case of facial expressions is to use computational models that can accomplish the same task as humans. Convolutional neural networks, which can recognize and categorize classes of objects, including faces, are a useful model of visual perception, and have been linked to activity in the human ventral visual pathway (Yamins et al., 2014). One such network is VGG-Face, which was trained to recognize the identity of faces from static images (Parkhi et al., 2015). Although this network was not trained to recognize emotion, a recent study found that VGG-Face learns representations that are sensitive to facial expression (Zhou et al., 2022). In addition, activations in VGG-Face can predict activity in the ventral visual pathway as measured by intracranial electroencephalography (Grossman et al., 2019). Given this evidence, VGG-Face is a valuable tool for investigating how the brain might process the visual inputs of facial expressions as it captures low level facial features associated with emotional expressions, as may be occurring in early stages of the ventral stream.

The current study aims to investigate whether person-specific or general mental models undergird emotion recognition by measuring the effect of prior experience with a target on inferences about their emotions. Using audiovisual clips from the TV show *Friends*, we experimentally manipulated the experience subjects had with several targets, while also capitalizing on subjects' past experience with the show. We determined whether these different levels of experience affected subjects' emotion judgments. To probe the role of facial expressions in familiar emotion recognition, we applied VGG-Face to the same faces that human subjects saw to determine whether its activations could predict human judgments. Given the

evidence for person-specific representations in social cognition, we hypothesized that more experience with a target would lead people to use a person-specific mental model of emotional expression when making emotion inferences. We predicted that subjects who had extensive past experience with the characters from *Friends* would use these individualized mental models, leading to a different pattern of emotion judgments compared with subjects who had less experience.

Methods

Participants

Thirty-eight participants (25 women) were recruited from Amazon Mechanical Turk. Participants were screened through CloudResearch and required to be working from the United States or Canada (Litman et al., 2017). They were adults between 25 and 65 years ($M = 43.9$, $SD = 10.1$, see Table S1 for complete demographics). All participants were compensated \$21 for their contribution to the study.

Determination of Expert vs. Naïve Status

Subjects were asked about their prior experience with *Friends* in order to determine the level of familiarity they had with the characters. Subjects who reported having seen a few episodes or less of the show were categorized as “naïve.” Those who had seen more than a few episodes (in most cases, many seasons) were categorized as “experts.”

Stimuli Selection

Stimulus clips were taken from the Multimodal Emotion Lines Dataset (Poria et al., 2018). This dataset contains video files, timing information, and transcriptions of the spoken lines for thousands of dialogues from the show *Friends*. For each of 4 characters (Joey, Phoebe,

Rachel, and Ross), exposure clips were selected showing dialogues between the given character and another character not in the target list. Together, the selection of exposure clips displayed each character expressing a range of emotions. Next, 35 test clips were selected for each of the 4 characters. Each of these was several seconds long ($M = 3.3$, $SD = 1.7$) and showed the character expressing an emotion; given their brief duration, these clips provided minimal context to support emotion inference.

Experimental Procedure

The experiment proceeded in four blocks (one for each character), with every block having exposure and test phases. Each character was assigned to a condition: audio, video, full, and text. Order and condition of the characters were counterbalanced across subjects. In the exposure phase, subjects were shown 8 clips of each character (described above). These clips varied in modality based on the condition: for the character assigned to the video condition, videos without audio were shown; for the audio condition, the audio was played while subjects viewed a blank screen; for the full condition, the unaltered audiovisual clip was shown; for the text condition, text from the dialogue spoken by the characters was displayed one line at a time, each with the same duration as in the clip. After each exposure phase, subjects viewed 35 short clips of the same character; these clips were a few seconds long and shown in unaltered audiovisual form (test phase). Subjects also viewed 31 short audiovisual clips of a character they had not been exposed to, as a control condition. After each clip in the test phase, subjects made a rating of the emotion they perceived the character to be feeling. They also made a rating of their own emotional state. They did so by clicking a location on a modified version of the Geneva Emotion Wheel (Scherer et al., 2013). This wheel has 20 emotion words arranged in a circle, with five concentric rows of dots that correspond to the emotion's intensity from low to high.

Subjects were trained on this response format prior to the experiment, and their understanding was checked by asking them to indicate a particular emotion/intensity. The time it took to make each rating was recorded.

Behavioral Data Analysis

Unsupervised Clustering

Unsupervised clustering was used to determine the structure of subjects' ratings of the characters' emotions. Specifically, Gaussian mixture modeling was used to group the data into clusters, each of which could be represented as a normally distributed density function (Fraley & Rafferty, 2002). Given the response format, which requires that subjects make ratings in 2D coordinate space on the Geneva Emotion Wheel, modeling groups of responses with two-dimensional Gaussians is appropriate. The *mclust* package in R was used for this approach (Scrucca et al., 2016). Clustering was performed for each group (expert, naïve) by condition (audio, video, full, text, control). To determine whether there were differences in precision between the groups, the average spread of the clusters was compared by computing the determinant of the covariance matrix for each cluster and averaging across all clusters. Next, to create a “ground truth” cluster solution, clustering was performed on the experts' ratings across all conditions. For each test clip that the subjects rated, the modal cluster assignment of that clip was used as the “correct” cluster assignment. 15 clusters were found as a stable solution, but one of these was not the modal assignment for any clip, leading to a ground truth mapping clips to 14 clusters.

Classification Using Cluster Solution

To determine whether the cluster solution that was generated from experts' ratings generalized within the expert group and across groups, classification was performed. Data from

half of the subjects in each group were designated as test data. With the *caret* package in R (Kuhn, 2008), a support vector machine with a radial basis function kernel was used for classification. Based on the spatial location of each subject's rating for each clip, it was classified according to the "ground truth" cluster solution. These predicted cluster assignments were compared to the "correct" assignment determined above to calculate classification accuracy. Accuracy was computed from the confusion matrix.

Reaction Times

We recorded how long it took for subjects to make an emotion rating for each clip. These reaction times were averaged within each condition (audio, video, text, control, and full audiovisual) and compared across conditions and expertise using a mixed ANOVA (within-subject variable = condition, between-subject variable = expert vs. naïve status). Additionally, reaction times were compared between trials in which subjects' ratings were classified as the correct cluster label (according to the consensus "ground truth" determined above) and trials in which the classification was incorrect. Within each subject, reaction times were averaged across all correct and all incorrect trials, and compared using a mixed ANOVA (within-subject variable = consensus, between-subject variable = expert vs. naïve status).

Convolutional Neural Network Analysis

Classification Using Cluster Solution

To generate inputs to VGG-Face, every frame of each test clip was passed through a face detection algorithm from MATLAB's Computer Vision Toolbox to select a cropped image of the character's face (MathWorks, 2021). These images were reviewed and images of non-target characters were removed. We next applied VGG-Face to these face-cropped images. Features from the units in the final convolutional layer (conv 5-3) that were previously found to be

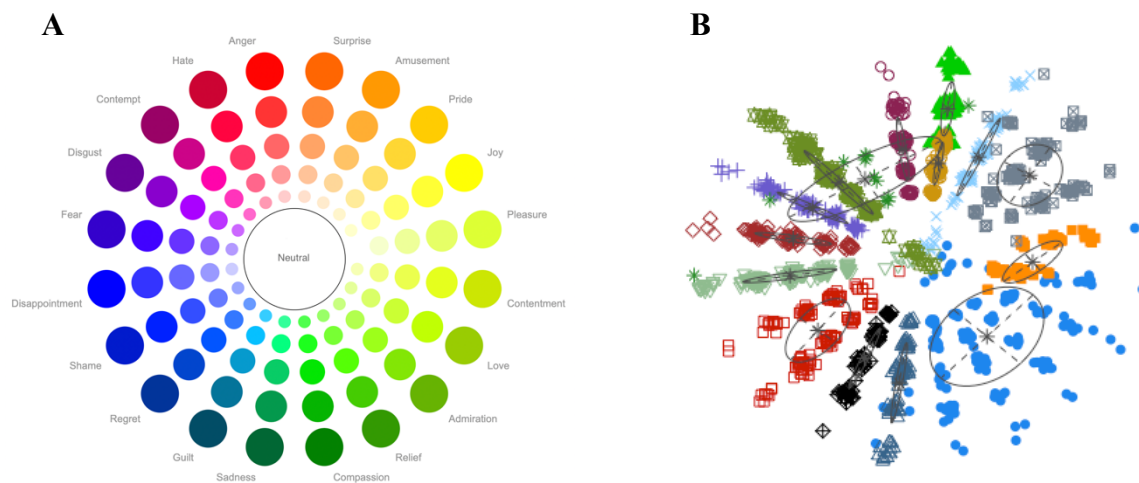
sensitive to expression were extracted (Zhou et al., 2022). These features were used in a partial least squares (PLS) regression model to predict the cluster membership of each frame. To test the accuracy of the model, hv-block cross-validation was used (Racine, 2000). In this approach, designed for handling the autocorrelation inherent in time-series data, some of the frames of each clip were designated for training the predictive model, while surrounding frames were removed from the test set. Classification accuracy was computed from a row-normalized confusion matrix.

Results

Unsupervised Clustering Reveals a Stable Expert Consensus in Perceived Emotion

A Gaussian mixture modeling cluster algorithm was used to group the coordinates of subjects' ratings into clusters represented by two-dimensional normal distributions. Across conditions, a 15-cluster solution emerged as a stable fit (see Figure S1).

Given that the 15-cluster solution reliably grouped ratings in each condition, 15 clusters were used to find a cluster solution for the expert ratings across all conditions (see Figure 1).



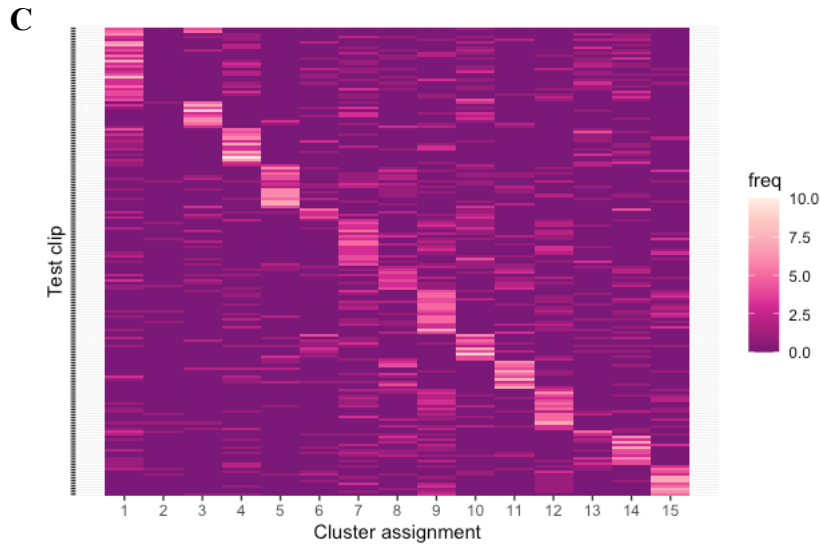


Figure 1. Emotion ratings can be grouped into 15 Gaussian clusters. **A)** The modified Geneva Emotion Wheel used in the study; subjects made ratings by clicking locations on the wheel. **B)** The Gaussian mixture model cluster solution for ratings made by the expert group. The ratings that fall into each cluster are represented by a different color and symbol. The ellipses are centered at the mean of each cluster, with length and width determined by the covariance of the Gaussian distribution. **C)** Heatmap showing how each test clip (rows) was assigned to each cluster (columns) based on the ratings of the expert group.

Classification into Consensus Solution Generalizes Across Expert and Naïve Groups

The classifier trained on the cluster assignments of half of the expert subjects was able to classify the ratings of both expert and naïve groups well above chance, at a similar level of accuracy (expert = 34.49%, naïve = 39.22%, see Figure 2). Given that the class distributions are unequal, we performed nonparametric tests to estimate chance performance by randomizing the assignment of stimuli to different clusters. These randomized models showed lower levels of performance (accuracy averaged across 1000 permutations for the expert group was 16.51%, naïve was 16.94%, $p = .001$). To characterize the nature of confusions between clusters, we

analyzed the classification errors between expert and naïve groups. If errors in classification are stemming from the same source of information (such as visual, auditory, and semantic properties of the stimuli), then they should be correlated. On the other hand, if the groups used different sources of information to make emotion inferences, then they should have different distributions of errors. Simple correlation of the confusions revealed a high level of consistency between groups ($r = .9064$), supporting the idea that a common source of information drove behavior. When each group's confusions were compared to confusions generated by randomly assigning stimuli to clusters, the correlation was moderate but significantly lower (expert $r = .5920$, naïve $r = .4602$, $p = .001$), indicating that a portion of the similarity in errors comes from unequal class distributions, while the rest can be attributed to similar emotion inference.

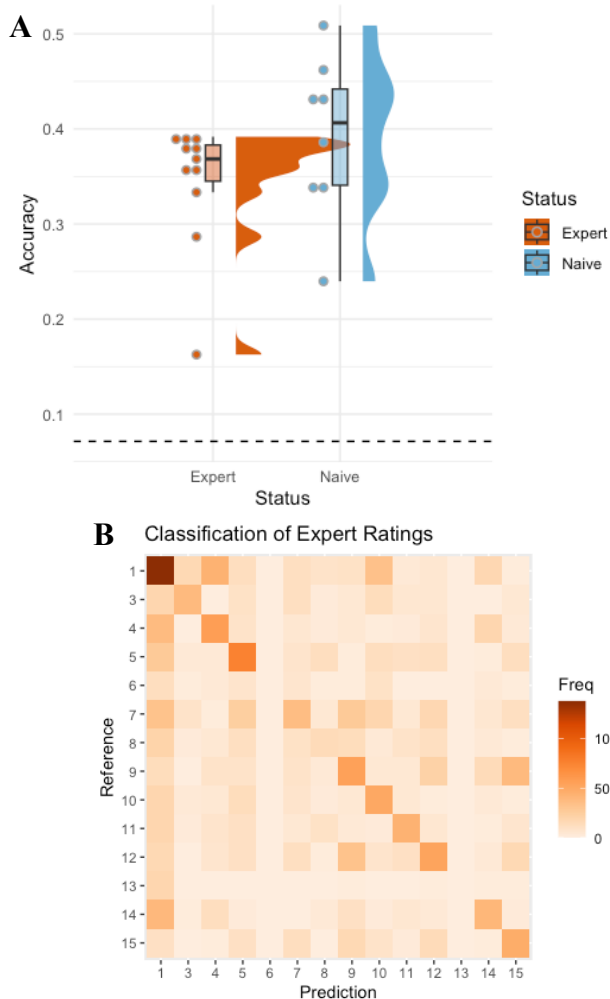


Figure 2. Classification generalizes across expert and naïve groups. **(A)** Accuracy of classification of each subject's ratings into the ground truth cluster solution, split by group. The dotted line shows the level of chance performance (1 over 14, or .0714). **(B and C)** Confusion matrices for the classification of expert and naïve ratings.

Reaction Time Differs By Consensus

Examining the timing of subjects' inferences could reveal differences in their efficiency when using a consensus model. Expert participants could be faster to make emotion inferences, and richer experience in the exposure phase of the study (as in the audiovisual condition) could lead to faster responses. Alternatively, participants could be faster when making ratings that match expert consensus, potentially indicating the more efficient use of a general mental model. On average, subjects took 6.739 seconds ($SD = 6.620$) to make an emotion rating in each trial. There was no significant difference in the time it took for subjects to rate the characters' emotions between conditions ($F(3,107.56) = 1.626, p = .188$) or groups ($F(1,36) = 3.604, p = .066$). However, there was a significant main effect of consensus: subjects were faster to make ratings that matched ground truth consensus than when they made ratings that did not, $F(1,17) = 23.46, p = .000152$.

Low-level Visual Information Does Not Predict Human Emotion Understanding/Perception

Because we found that expert and naïve participants showed a similar structure of emotion inference, we wanted to examine the physical properties of the stimuli. To test the role of facial expressions, a prominent channel for conveying emotion, we used expression-sensitive units from the deep convolutional network VGG-Face to classify clips into the cluster solution determined from the human ratings. Using *h*v-block cross-validation to sample a subset of video frames from clips, we found the model predicted human judgments with an accuracy of 36.76%; however, permutation tests revealed that this did not meaningfully differ from classification into categories whose labels were randomly shuffled ($p = .45$). Performing cross-validation by holding out entire test clips also resulted in low levels of performance (6.18% accuracy). Together, these findings suggest that it is unlikely that participants were using invariant

components of facial expression to make emotion judgments, and that other perceptual or conceptual information informed their decisions.

Discussion

Humans make emotion inferences from a variety of signals, and do so more accurately for those they know, but it is unclear how the process of mapping expressions to emotions might differ in that case. Among the range of possibilities for how identity information affects expression processing, there are two ends of the spectrum: 1) they remain relatively independent, with a general mental model implemented to interpret expression across all targets, and 2) they interact to produce person-specific mental models of expression. The current study interrogated this dichotomy by asking subjects with varying levels of experience with particular targets to make inferences about their emotions. Using unsupervised clustering to examine the structure of subjects' ratings, we found no differences in the precision or mapping of emotion judgments by experimental condition or group. In fact, a classifier trained on the cluster solution of the expert group's ratings accurately classified both held-out experts' ratings and naïve subjects' ratings, indicating a systematic mapping of emotional expression to cluster (a "consensus") regardless of the subject's level of experience. Reaction time also did not differ by condition or group, but was significantly longer when subjects made ratings that did not fit with the consensus model. These results suggest that a general mental model of emotional expressions undergirded subjects' emotion inferences in this study.

There are several possible interpretations of these results in terms of their implications for how experience impacts emotion understanding. The first and most extreme is that a general mental model is the only model involved in emotion inference, regardless of who is making the

expression. This interpretation stands in contrast to prior research showing that empathic accuracy is higher for known compared to unknown targets (Stinson & Ickes, 1992; Thomas & Fletcher, 2003). While our finding that consensus generalized across expert and naïve participants is compatible with this interpretation, given prior research it seems unlikely that a general mental model is sufficient in every case. The second interpretation is that while person-specific mental models may be implemented in the brain to guide inference in ambiguous situations, the general mental model is sufficient in many situations, including the emotion judgments made in this study. Given the fact that the emotional expressions subjects saw were presented with rich audiovisual information in the context of an acted scene, it is possible that expressions were clear enough in this study to use a general mental model for inference. In other more ambiguous situations, person-specific mental models might contribute to emotion understanding. The third interpretation is that the naïve subjects in this study were not naïve enough. Indeed, some reported seeing one to two episodes of *Friends*, which exposed them to the characters and their expressions outside of the experimental conditions. Given the popularity of *Friends*, others in the naïve group might have been exposed to the show and its characters in other ways, including through the internet. Additionally, even though the actors' portrayal of these characters presumably differs from their emotional expressivity in other roles, it is possible that knowledge gained from other exposure to the actors provided naïve subjects with some expertise. If many subjects in the naïve group had some meaningful level of exposure, then they might already have developed person-specific mental models for these targets, rendering the structure of their emotion judgments similar to the experts'.

Whichever of these possibilities is the case, it is clear that a consensus emerged across subjects. Clusters derived from half of the experts' ratings via an unsupervised approach were

used to train a classifier, and clips were reliably mapped to the same clusters in both expert and naïve groups. Not only was accuracy of classification similar across groups, but the errors made were highly related. In other words, for a given emotional expression, expert and naïve subjects rated it similarly in cluster-space. One problem in studying emotion recognition using complex, naturalistic stimuli is the lack of a “ground truth” (Zaki & Ochsner, 2012). When moving beyond prototypical emotional displays reflecting a small number of categories, and without the ability to ask the target directly, it is difficult to determine what an “accurate” emotion inference is. The approach used in this study—looking across people to see if emotion judgments converge on some consensus understanding—is a useful proxy for ground truth. Even with variable, multimodal displays of emotion such as those in this study, similar emotions were recognized for the same expression across subjects. These results support the approach of using consensus across people as an operationalization of ground truth.

Examining the pattern of subjects’ reaction times revealed that the temporal dynamics of emotion inference differed between instances when subjects rated with vs. against consensus. When deciding what emotion a character was feeling, subjects took significantly longer to make choices that did not match the consensus. Perhaps making an emotion inference that does not map on to the consensus model reflects deliberation occurring in more ambiguous situations; this could be the case when there is competing information from different modalities (such as a smile paired with an angry tone of voice). However, results showing differences in reaction time alone should be interpreted cautiously; a controlled experiment manipulating multimodal congruence/ambiguity is necessary to make strong claims about the psychological processes at play.

In some situations, visual features such as those captured by VGG-Face might be able to be used to rapidly and generally detect emotional expressions. However, when VGG-Face was applied to the same stimuli that subjects saw, a classification model based on its activations in expression-selective units was not able to meaningfully map facial expressions to the clusters derived from subjects' ratings. In effect, VGG-Face extracted static visual features from the expressions made by the characters, and these were not sufficient to explain human inference. It is likely that dynamic qualities of facial expressions are important for human emotion understanding, and a neural network model incorporating dynamics may better predict human judgments. Of course, faces were not the only sources of meaningful signal that subjects saw; information from other modalities which was not captured by VGG-Face likely contributed to their understanding of the characters' emotions. In addition, there are differences between our study and past approaches that demonstrated expression-selectivity in VGG-Face. When Zhou and colleagues (2022) found units in VGG-Face that were selective for emotional expression, they used images of faces that were well-lit, looking directly ahead, and fully cropped of all background. In contrast, the faces used from test clip frames had variable lighting and positioning and contained some background around the edge of the target's face. These differences could have introduced noise that reduced the ability of VGG-Face to reliably pick up on emotional expression. Although VGG-Face could not explain human behavior in this study, recent views in computational cognitive science emphasize the value of using neural networks to map out a space of hypotheses, rather than searching for one all-encompassing computational explanation (Golan et al., 2023). From this standpoint, failures of models provide useful insight into the necessary and sufficient components of cognitive processes.

A strength of this study is its use of multimodal, naturalistic stimuli in studying emotion recognition. Compared to many studies using images of posed, static expressions, watching audiovisual clips is much closer to real-world emotion understanding. Subjects conveyed their emotion inferences not using a limited set of five to six emotion categories, as is the case in many studies of emotion recognition, but instead with a response format that allowed them to select between 20 categories and simultaneously indicate intensity. In addition, this study made use of an experimental manipulation (varying the modality of exposure) and subjects' past experience (expert vs. naïve) to tap into multiple degrees of exposure. Along with these strengths, there are several limitations of this research. As noted above, the naïve group may have had enough prior experience with the targets to invalidate their categorization as "naïve." Additionally, data were collected online using Amazon's Mechanical Turk platform, leading to a non-representative sample and the possibility of inattention/bots, although screening of participants assuages some of these concerns (Buhrmester et al., 2018). Finally, features from VGG-Face were not able to accurately predict human emotion judgments, limiting our ability to use it as a tool to explain human behavior.

To address these limitations, future research directions are threefold. First, refining the approach for how VGG-Face is applied to the target stimuli could allow it to better predict human behavior. This could be done by using another model as a preliminary filter to identify clearer facial displays given potential issues with noise. Second, other modalities should be explored to investigate how they contribute to emotion judgments. Semantic information from language is a particularly suitable target, given past studies showing the importance of verbal content for empathic accuracy (Gesn & Ickes, 1999; Jospe et al., 2020). Natural language processing models could be applied to find semantic features in the dialogue which could then be

used to predict subjects' emotion judgments. Third, a similar experimental approach could be taken with a more controlled manipulation of exposure. Subjects could be presented with targets they had never encountered before and could be assigned to experimental conditions with varying amounts of exposure. This would allow for a stricter test of the effect of experience on emotion understanding.

This study examined the impact of experience on emotion understanding. The results demonstrate that a shared consensus guided subjects' ratings of targets' emotions, regardless of their past experience with the targets. Although there are outstanding limitations that need to be addressed, this study suggests that at least in some situations, a general mental model of emotional expressions is sufficient for making inferences about others' emotions. Future work is necessary to more tightly control the manipulation of experience, and to tease apart the contributions of different modalities to the process of emotion understanding.

References

- Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on psychological science*, 13(2), 149-154.
- Calder, A. J. (2011). Does facial identity and facial expression recognition involve separate visual routes. *The Oxford handbook of face perception*, 427-448.
- Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8), 641-651.
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in cognitive sciences*, 11(12), 535-543.
- Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Current biology*, 19(12), 1028-1033.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458), 611-631.
- Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology*, 77(4), 746.
- Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *Neuroimage*, 22(4), 1628-1635.
- Golan, T., Taylor, J., Schütt, H. H., Peters, B., Sommers, R. P., Seeliger, K., Kriegeskorte, N. (2023). Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. *PsyArXiv*, <https://doi.org/10.31234/osf.io/tr7gx>

- Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., ... & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature communications*, 10(1), 4934.
- Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion*, 7(2), 438.
- Harry, B. B., Williams, M., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in human neuroscience*, 7, 692.
- Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological psychiatry*, 51(1), 59-67.
- Hohwy, J. (2013). *The predictive mind*. OUP Oxford.
- Holt-Lunstad J., Smith T.B., & Layton J.B. (2010). Social relationships and mortality risk: A Meta-analytic review. *PLoS Med* 7(7): e1000316.
- Jospe, K., Genzer, S., klein Selle, N., Ong, D., Zaki, J., & Perry, A. (2020). The contribution of linguistic and visual cues to physiological synchrony and empathic accuracy. *Cortex*, 132, 296-308.
- Kragel, P. A., & LaBar, K. S. (2016). Somatosensory representations link the perception of emotional expressions and sensory experience. *eneuro*, 3(2).
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28, 1-26.
- Leibenluft, E., Gobbini, M. I., Harrison, T., & Haxby, J. V. (2004). Mothers' neural activation in response to pictures of their children and other children. *Biological psychiatry*, 56(4), 225-232.

- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433-442.
- The MathWorks Inc. (2021). Computer Vision Toolbox (R2021b), Natick, Massachusetts: The MathWorks Inc. <https://www.mathworks.com>
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.
- Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of neuroscience*, 30(30), 10127-10134.
- Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- Racine, J. (2000). Consistent cross-validators model-selection for dependent data: hv-block cross-validation. *Journal of econometrics*, 99(1), 39-61.
- Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1), 79-87.
- Scherer, K.R., Shuman, V., Fontaine, J.R.J., & Soriano, C. (2013). The GRID meets the Wheel: Assessing emotional feeling via self-report. In Johnny R.J. Fontaine, Klaus R. Scherer & C. Soriano (Eds.), *Components of Emotional Meaning: A sourcebook* (pp. 281-298). Oxford: Oxford University Press.
- Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, *The R Journal*, 8/1, pp. 205-233. <https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf>

- Sened, H., Lavidor, M., Lazarus, G., Bar-Kalifa, E., Rafaeli, E., & Ickes, W. (2017). Empathic accuracy and relationship satisfaction: A meta-analytic review. *Journal of Family Psychology*, 31(6), 742.
- Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of personality and social psychology*, 62(5), 787.
- Thomas, G., & Fletcher, G. J. O. (2003). Mind-reading accuracy in intimate relationships: Assessing the roles of the relationship, the target, and the judge. *Journal of Personality and Social Psychology*, 85(6), 1079–1094.
- Tsantani, M., Kriegeskorte, N., McGettigan, C., & Garrido, L. (2019). Faces and voices in the brain: a modality-general person-identity representation in superior temporal sulcus. *NeuroImage*, 201, 116004.
- Tsuchiya, N., Kawasaki, H., Oya, H., Howard III, M. A., & Adolphs, R. (2008). Decoding face information in time, frequency and space from direct intracranial recordings of the human brain. *PloS one*, 3(12), e3892.
- Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face–voice emotional integration. *Journal of Neuroscience*, 34(20), 6813-6821.
- Welborn, B. L., & Lieberman, M. D. (2015). Person-specific theory of mind in medial pFC. *Journal of Cognitive Neuroscience*, 27(1), 1-12.
- Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of neurophysiology*, 92(3), 1830-1839.

- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619-8624.
- Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nature neuroscience*, *15*(5), 675-680.
- Zhao, Z., Thornton, M. A., & Tamir, D. I. (2020). Accurate emotion prediction in dyads and groups and its potential social benefits. *Emotion*.
- Zhou, L., Yang, A., Meng, M., & Zhou, K. (2022). Emerged human-like facial expression representation in a deep convolutional neural network. *Science advances*, *8*(12), eabj4383.

Supplemental Materials

Table S1. Demographics

| Measure | <i>n</i> | % |
|---------------------------|----------|------|
| Gender | | |
| Female | 25 | 65.8 |
| Male | 13 | 34.2 |
| Race | | |
| White | 30 | 78.9 |
| Black | 2 | 5.3 |
| Asian | 5 | 13.2 |
| Native American | 1 | 2.6 |
| Highest educational level | | |
| Less than high school | 1 | 2.6 |
| High school/some college | 18 | 47.4 |
| College degree | 17 | 44.7 |
| Postgraduate degree | 2 | 5.3 |

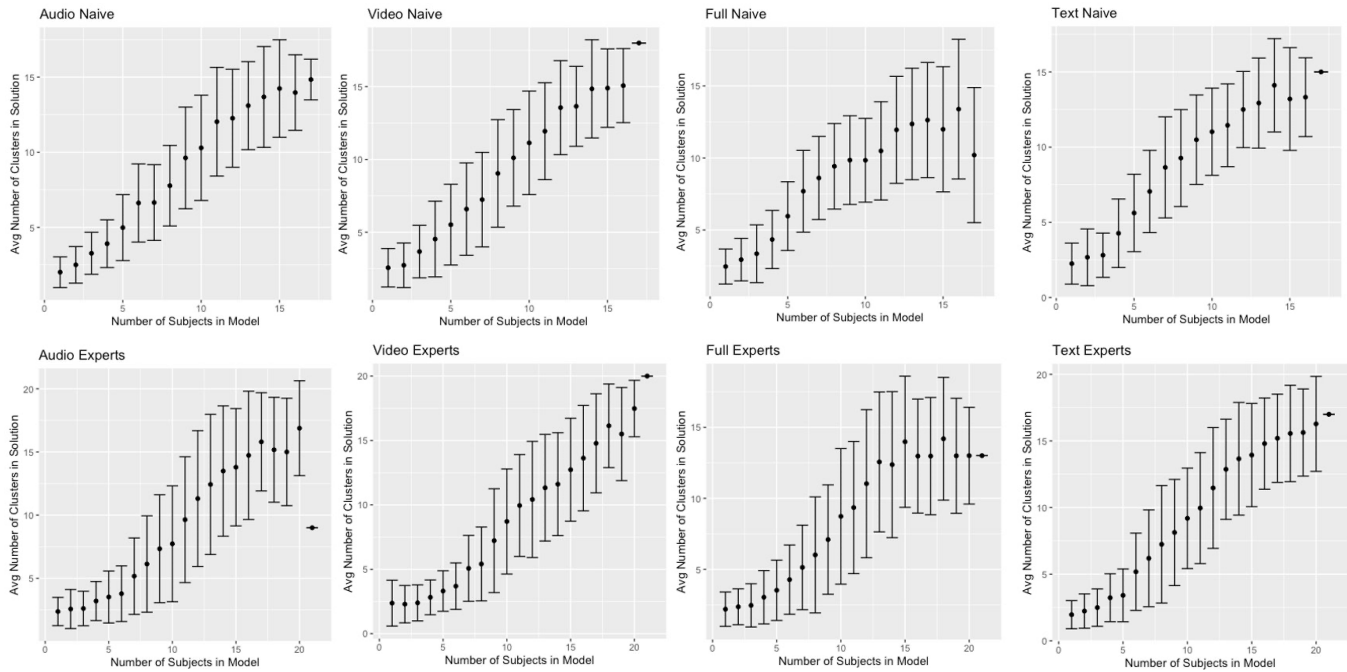


Figure S1. Unsupervised clustering of each condition by group led to cluster solutions that stabilized around 15 clusters. Each panel shows the average number of clusters (± 1 SD) in the clustering solution derived from a random subsample of subjects over 100 iterations. From left to right, panels are arranged showing the audio, video, full audiovisual, and text conditions. Top row = naïve group, bottom row = expert group.

An analysis comparing the average precision of clusters revealed no differences between experimental conditions or between expert and naïve groups.