2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)

2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) | 978-1-7281-3079-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/FG47880.2020.00123

Emotion or expressivity? An automated analysis of nonverbal perception in a social dilemma

Su Lei, Kalin Stefanov, Jonathan Gratch
Institute for Creative Technologies, University of Southern California, Los Angeles, USA

***Abstract*— An extensive body of research has examined how specific emotional expressions shape social perceptions and social decisions, yet recent scholarship in emotion research has raised questions about the validity of emotion as a construct. In this article, we contrast the value of measuring emotional expressions with the more general construct of *expressivity* (in the sense of conveying a thought or emotion through any nonverbal behavior) and develop models that can automatically extract perceived expressivity from videos. Although less extensive, a solid body of research has shown expressivity to be an important element when studying interpersonal perception, particularly in psychiatric contexts. Here we examine the role expressivity plays in predicting social perceptions and decisions in the context of a social dilemma. We show that perceivers use more than facial expressions when making judgments of expressivity and see these expressions as conveying thoughts as well as emotions (although facial expressions and emotional attributions explain most of the variance in these judgments). We next show that expressivity can be predicted with high accuracy using Lasso and random forests. Our analysis shows that features related to motion dynamics are particularly important for modeling these judgments. We also show that learned models of expressivity have value in recognizing important aspects of a social situation. First, we revisit a previously published finding which showed that smile intensity was associated with the unexpectedness of outcomes in social dilemmas; instead, we show that expressivity is a better predictor (and explanation) of this finding. Second, we provide preliminary evidence that expressivity is useful for identifying "moments of interest" in a video sequence.**

I. Introduction

Inspired by the pioneering work of Paul Ekman and early appraisal theories [1], much of the work in affective computing follows a "standard model" which argues that: (1) events of personal significance to an individual are appraised and trigger an emotional response and (2) this response is reflected in external emotional signals, especially facial expressions, as a window into the affective state [2] and (3) these expressions influence the behavior of perceivers (e.g., through contagion or inferences about the senders' affective state) [3] [4]. In line with these views, many studies have collected data of social interactions, examined facial expressions, and made predictions about significant events. Vinkemeier et al. [5] tried to predict poker folds from face reactions to events in a poker game. Hoegen et al. [6] tried to predict cooperative or noncooperative responses based on facial reactions to events in a social dilemma. Mussel et al. [7] found that offers in an ultimatum game were more often accepted if the proposer smiled and less often accepted if the proposer showed angry facial expression.

However, the predominant view in emotion research today is that the "standard model" is incorrect, or at least requires significant qualification. For example, Jack et al. [8] and Du et al. [9] argue that emotions are neither basic or universal. Others see emotional expressions as communicative acts that shape social encounters [10]. Thus, they are not necessarily a reflection of the underlying emotional state [11], and share much with other communicative acts (words, gestures) [12].

If the "standard model" is incorrect, it provides opportunities to approach old questions from a new perspective. In this article, we consider several research questions using a corpus of spontaneous reactions to personally significant events. (1) Are facial expressions the best indicators that something personally significant has just happened to an individual, or are other nonverbal behaviors equally diagnostic? (2) What behaviors do perceivers use to make inferences that something significant has occurred to this individual? Are facial expressions predominant, or are other behaviors equally important? (3) What temporal aspects of these behaviors are crucial for supporting these inferences [13][14]? (4) Do perceivers see these expressions as emotional, or do they feel they communicate thoughts? In addressing these questions, we step back from a focus on facial expressions and consider a more general construct of expressivity. (5) Could the perception of expressivity lead to better social inferences and predictions? Specifically, we address these questions in the context of a social dilemma. From Amazon Mechanical Turk, we crowdsource judgments of the extent to which a person is reacting to significant events in an iterated prisoner's dilemma and how expressive they are. We collect information on what behaviors the observers used to make these judgments (face, head, body, hands). Then we build an automatic recognizer that predicts observers' judgments on expressivity. Our findings suggest that the observers made inferences about the senders' expressivity based on the temporal dynamics of nonverbal behaviors, and that they attend mostly to people's facial expressions. We also show that expressivity predicted by our model outperforms specific facial expressions (smiles) in predicting significant events.

II. Related Work

A. Expressivity as a Construct

Although emotion and affective computing research has emphasized the importance of specific facial expressions, researchers of nonverbal communication (and, indeed, the face and gesture community) have taken a broader view of nonverbal signals. Within this broader tradition, nonverbal expressivity (i.e., the presence and strength of behaviors that convey some thought or emotion) has been shown to have a

978-1-7281-3079-8/20/$31.00 ©2020 IEEE
DOI 10.1109/FG47880.2020.00123

profound impact on interpersonal perception and outcomes [15]. For example, work by Burgoon et al. [16] and Berneiri et al. [17] characterized expressivity in terms of presence and dynamics of facial movements, gestures, and posture, and they found that expressivity was a primary factor in the establishment of rapport between speakers. Even when this work has emphasized emotional behaviors, but has considered the presence of emotional expressions *in toto*, rather than examining the presence of specific expressions. Similar to expressivity, past work from the face and gesture community has also taken a holistic approach, Hernandez et al. [18] built a model with face and head gestures to automatically measure the engagement level of TV viewers. Admittedly, not all nonverbal behaviors are necessarily expressing a specific emotion. Therefore we could potentially benefit from learning them with a more generalized construct.

Expressivity has been examined from various perspectives. For example, Boone and Buck [19] considered expressivity "as the accuracy with which an individual displays or communicates his or her emotions." From an evolutionary perspective under the context of a social dilemma, they argued that emotional expressivity signals trustworthiness and serves as a marker for cooperative behaviors. In clinical research, expressivity also plays a critical part in studying affective features and the disorders of social interactions among psychiatric patients. A severe reduction in facial expressivity or irregularity in nonverbal production, is associated with conditions such as schizophrenia, depression, autism, and Parkinson's Disease. Evidence had shown that schizophrenia patients displayed atypical expressions and were less facially expressive than controls [20], even so when they experienced as much emotion [21]. Girad et al. [22] found that when symptoms were severe, patients with depression regulated interpersonal distance by displaying more facial action units associated with negative emotion and less associated with positive emotion. A recent meta-analysis [23] suggested that facial expressions of people with autism are atypical. Georgescu et al. [24] advocated to use virtual characters to assess and train individuals with high-functioning autism, and further help them improve social skills. Buck et al. [25] proposed a technique to study the emotional expression and communication style of behaviorally disordered children and schizophrenic patients and their family members. Mounting evidence has suggested that advance in automatic recognizing and understanding of expressivity could help us better study social interaction and help develop diagnostic and treatment tools for clinical assessment.

B. Measuring and Predicting Expressivity

Traditionally expressivity was measured either by self-assessment (sender) or by experts conducting time-consuming manual annotations (observer). Tickle [26] developed a rating protocol for the observers to measure expressive behavior for patients with Parkinson's Disease. Kring et al. [27], Gross and John [28] built two well-validated self-assessment tools to measure the extent to which people consider themselves "outwardly exhibit emotions" or "reveal feelings." The questionnaires ask people to rate themselves on questions such as "I display my emotions to other people" or "No matter how nervous or upset I am, I tend to keep a calm exterior." These tools assess the expressivity of oneself as a personality trait, and they emphasize that the definition of emotional expressivity is not limited to a specific emotion (though [28] provides subscales of positive expressivity and negative expressivity), or limited to a specific modality/channel of expression. Among all modalities, facial expressivity has been studied most extensively. Along with the advancement in computer vision, researchers can integrate automatic facial expression recognition tools to gain insights into facial expressivity measurement. Neubauer et al. [29] used tracked facial expressions to represent facial expressivity directly; Wu et al. [30] developed a more nuanced arithmetic calculation based on Tickle's protocol.

To distinguish current work from self-report of the senders' emotional expressivity, we focus entirely on perceived expressivity from the perspective of the observers. More importantly, we aim to build an automatic predictor of perceived expressivity. There is little work done similarly. To study patients with Parkinson's, Joshi et al. [31] acquired ground truth ratings based on Tickle's protocol, and built machine learning models to predict expressivity from automatically tracked facial features. More recently, Lin et al. [32] investigated the perceived expressiveness of senders participating in different emotional tasks, and they found nonverbal features associated with perceived expressivity differed by emotional contexts. Though in our context of a social dilemma, senders might experience different emotions. We do not draw such distinction, and instead, focus our investigation on how different modalities, and their temporal dynamics, interact to determine perceived expressivity.

III. Expressivity Corpus

To examine the importance of expressivity as a construct, we identified an existing large corpus of individuals spontaneously reacting to personally significant events, and recruited a large panel of crowd workers to annotate the perceived expressivity of these reactions.

A. Iterated Prisoner's Dilemma Corpus

The iterated prisoner's dilemma is a standard social dilemma that is often used to study emotional reactions and the role facial expressions play in shaping joint decisions (e.g., [33], [3]). To study expressivity, we used the USC Iterated Prisoner's Dilemma (IPD) Corpus containing more than 6000 spontaneous nonverbal reactions to decisions in this game [34]. The IPD Corpus contains videos of 716 individuals (51% female, age 18-65) playing a web-based version IPD modeled after the UK TV show Golden Balls. The study was approved by the Institutional Review Board (IRB) of University of Southern California. Participants were recruited from Craigslist and played ten rounds with the same opponent that they were randomly paired with. In each round, players simultaneously chose either to split or to steal, and received points based on a playoff matrix. To guarantee

nonverbal interaction, players could not talk to each other and could only see each other's body above the chest through a webcam (Fig.1). To ensure players were engaged and motivated to perform, they were compensated based on their decisions in the game. In addition to a $25 participation fee, they received lottery tickets based on the points they earned, and these were entered into seven $100 lotteries. Since our

Fig. 1. Each player gets 5 tickets if both choose to cooperate (*split*), 1 ticket if both choose to defect (*steal*). When the player chooses to cooperate (*split*) and their opponent chooses to defect (*steal*), the player gets nothing while their opponents gets 10 tickets.

goal is to study perceived expressivity, we focus on the most eventful video segments (seven seconds) where players show their reactions to the game result following each round.

B. Annotation Task

Crowd workers rated randomly selected videos, each video depicting a single player of the moment they learned the outcome of one of the rounds of their game. Annotators were told these videos showed players' reactions within a two-person game, but were not told the specific outcome of the joint decision. Workers were asked to rate the individual's expressivity. As our goal was to understand the specific behaviors that workers' used to make these judgments, crowd workers were asked which component of the body contributed to their judgment. As another goal was to understand if perceivers view these behaviors as strictly emotional, we asked crowd workers to indicate the extent to which the reaction conveyed a thought or emotion. Finally, we asked crowd workers to provide a brief description of what they felt was expressed (see questions in Section III-B.2).

1) Annotators: We recruited 274 crowd workers from Amazon Mechanical Turk to rate a subset of 1000 video segments from IPD. Videos from participants who declined to share their video recordings were excluded from the rating task. Each crowd worker rated 20 randomly presented videos. They were allowed to watch each video as many times as they wanted to.

2) Inter-rater Reliability: Intraclass correlation coefficient (ICC) was calculated to assess if crowd workers were able to provide consistent responses. Each video was rated by a different group of (randomly selected) observers, and we combined these into a single mean rating for each video. Thus, a one-way random ICC(1,k) was used [35] to assess agreement. Due to the limitation of the randomization process of our survey platform, we received an uneven number of ratings for each video. Most videos received 5 or 6 ratings, N(5)=513, and N(6)=437. A few videos received 4 or 7 ratings, N(4)=19, and N(7)=31. To calculate ICC(1,k), we chose k=5. Videos received only 4 ratings were treated as one rater missing, and we randomly sampled 5 ratings for videos received more than 5 ratings.

For each video, we asked the observers eight 7-point Likert questions and one free form text question. For all Likert items, we achieved overall ICC=0.76, with a 95% confidence interval (CI) from 0.75 to 0.77. We also report ICC for each Likert question as follows:

1) How strongly is the person reacting to the event? (*reaction*, ICC=0.80, 95% CI [0.78, 0.82])
2) How expressive was the person? (*expressivity*, 0.80, 95% CI [0.78, 0.82])
3) What part of the body conveyed these impressions?
   - facial expressions (*face*, 0.77, 95% CI [0.75, 0.79])
   - head movements (*head*, 0.70, 95% CI [0.67, 0.73])
   - posture movements (*posture*, 0.58, 95% CI [0.54, 0.62])
   - hand or arm movements (*hand*, 0.62, 95% CI [0.58, 0.66])
4) To what extent does the person seem to be expressing...
   - emotion (*emotion*, 0.78, 95% CI [0.75, 0.80])
   - a thought or concept other than emotion (*thought*, 0.27, 95% CI [0.19, 0.34])
5) In as few words as possible, what thought or emotion is being expressed?

Inter-rater reliabilities are all within reasonable range except for the *thought* item. In other words, the observers agreed on the extent to which people were expressing emotion, but had different opinions on whether a thought or concept other than emotion was being expressed. To examine in greater detail, we did further analysis with the texts the observers described in Section III-B.4.

3) Analysis of ratings: We perform statistical analysis to understand how the observers made inferences from all modalities. As shown in Fig.2, we see a very high correlation between *reaction* and *expressivity* (Pearson's r=0.97). As observers considered these two questions almost identical, we collapse them and use the mean of the two items as an *expressivity score* for the ground truth of our predictive models in the next section. *Expressivity* and *reaction* correlates the most with *face* (r=0.92) among other modalities. To examine how much these modalities contribute in conveying the impressions to the observers, we fit a multiple linear regression with standardized *face*, *head*, *posture* and *hand* to explain composited *expressivity score*. The model was highly significant (F=6.03, p<.0001), and all modalities combined explain 90.7% (R^2) of variance in *expressivity score*. Not surprisingly, *face* contributed the most (β=0.73, p<.0001), *head* (β=0.15, p<.0001) and *posture* (β=0.15, p<.0001) contributed almost the same but much less than *face*. Last, *Hand* had a very small but significant contribution, β=0.03, p=0.02.

We can also see in Fig.2, *emotion* is highly correlated with *reaction* and *expressivity* (r=0.94). It suggests that the more expressive the observers considered the person was, the

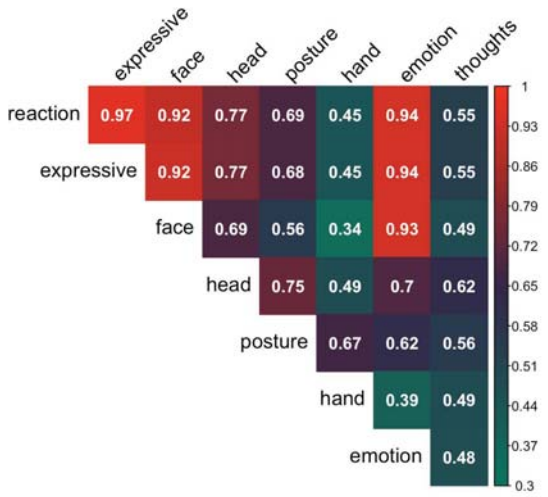

Fig. 2. Pairwise correlations (Pearson's r) among rating tasks. We choose Pearson's r after visual inspection of each rating item's histogram. All items fairly follow a normal distribution except head, posture and hand.

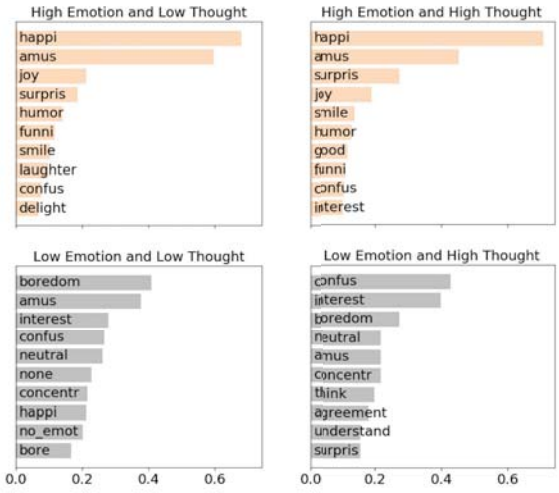

Fig. 3. Representative words (stemmed) described by observers.

more confident they were that the person was expressing an emotion. *Emotion*'s high correlation with *face* (p=0.93) tells us that the more the observers perceive expressivity from people's face, the more likely that the observers consider the person to be expressing an emotion.

4) Analysis of open-ended responses: We used the open-ended descriptions of reactions to get a sense of what was being conveyed. As one of our research questions was to understand if nonverbal behaviors conveyed more than emotion, we grouped videos into four categories based on the observers' ratings for the *emotion* and *thought* items. Specifically, for *emotion* item, we performed median splits to divide videos into "high emotion" or "low emotion." Similarly, we used the *thought* item to divide videos into "high thought" or "low thought." For each video, we concatenated the text descriptions provided by each crowd worker, removed stop words, and stemmed the rest. Then we calculated term frequency-inverse document frequency (TFIDF) for the combined texts. TFIDF is a common weighting scheme in text analysis [36]. A higher value means the word appears more often in a certain group after taking into consideration that this word might appear more often in general among all groups. Words with the highest ten TFIDF scores for each group are shown in Fig.3. Most noticeably, these descriptions differ by the extent to which the observers perceived people were expressing emotion, and there is no such difference on *thought* dimension. Videos perceived as highly expressing emotion are associated with words such as "happy," "amusement," "joy," and "surprise," otherwise associated with words such as "boredom," "concentration," "interest," "confusion," and "neutral."

IV. Predicting Perceived Expressivity

In this section, we introduce how we build models that can predict perceived expressivity from features automatically extracted from videos.

A. Feature Extraction

Here we describe how visual features are automatically extracted, and (in the next section) grouped to capture modality of "face," "head," and "posture." We do not have proper measures for "hand," and we also think it is reasonable to disregard it due to its limited contribution to observers' perception.

1) Facial Expressions: A commercial software based on CERT [37] is used to track the intensity of 20 frame by frame Action Units (AUs) [38]. We also construct six *facial factors* based on [39]. In this recent work, they performed factor analysis on AU features and discovered six psychologically meaningful factors; the factors are Enjoyment Smile (F1), Eyebrows Up (F2), Open Mouth (F3), Mouth Tightening (F4), Eye Tightening (F5) and Mouth Frown (F6). For each AU and *facial factor*, velocity is also calculated to describe the velocity of change in intensity. Then we compute the average, standard deviation, and max for each of the signals. Thus each signal is represented by six features in total. Past work has suggested that AU's activation rate varies. It is also known that automatic tracking software's detection accuracy varies by AU as well. Indeed what we have seen in IPD supports these views. Taking these views into consideration, we create composite facial signal features from AUs with at least a minimal activation rate. The activation rate of AU was calculated across the entire IPD corpus. Ten of them were activated at least 15% of all frames. Six composite features were created by summing up values across ten activated AUs.

2) Facial movements: ZFace [40] was used to track 49 facial landmarks. Similar to [41], for each one of the facial landmarks, we calculated the current frame's displacement from the mean position of the individual. Then we calculated the velocity of displacement. Principle component analysis was performed on both displacement features and velocity of displacement features to reduce 49 facial landmarks to two

dimensions (94%, 86% variance was preserved respectively). Finally, we calculated average, standard deviation, and max values for landmarks' displacement and velocity as landmark features.

3) Head movements and gestures: ZFace was also used to track head orientations in three directions (pitch, yaw, and roll). Similar to processing facial landmarks, for each direction, displacement from the individual's mean and the velocity of displacement were calculated. Then we calculated average, standard deviation, and max values for each direction as head movement features. In addition, we used the head gesture detector described in [42] to get frame by frame head nod and head shake binary prediction. Then for each video, we counted the number of head nods and head shakes as head gesture features.

4) Optical Flow: Optical flow captures movement between consecutive frames caused by either motion of image objects in the frame or by moving the camera. In IPD, videos were recorded from a webcam attached to a desk monitor. Thus motion captured by optical flow can only be contributed by movements of player.[1] As seen in Fig. 1, the camera captures a player's body part above the chest, so optical flow could potentially provide us with additional information of posture movement. We computed dense optical flow using OpenCV's implementation of Gunner Farneback's algorithm [43]. For each frame, we took the sum of flow magnitude across all pixels. Then we calculated average, standard deviation, and max values as our flow features.

B. Modalities, Feature Sets, and Label

We group the extracted features by modality.

1) Baseline: In past work, the sum across averaged AUs or the count of the six basic emotion label occurrences (joy, surprise, sadness, anger, disgust, and fear) were often used directly as an estimate of total expressivity [29] [44]. In line with this, we use 19 averaged AUs[2] as our baseline feature sets.

2) Face Feature Set: All the facial expression features are included in the *Face Feature Set*; we have 156 features in total.

3) Head Feature Set: All the head movement and gesture features are included in the *Head Feature Set*; we have 26 features in total.

4) Posture Feature Set: Though we do not have features that are directly tracking body postures, during our exploratory analysis we found that facial landmark features and optical flow features best described posture movements. Heuristically, the positions of landmarks are very likely to move while one changes their posture. Optical flow features capture visual movements between consecutive frames; thus posture movements should be captured among others. We have a total of 21 posture features.

[1]We noticed that in some videos, the experimenter's movements in the background could also be captured by the camera. One limitation of this method is that we cannot filter this noise.

[2][37] outputs AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, AU28 and AU43. All AUs except AU43 (Eye Closed) were included in our analysis.

5) Dynamics: Within each modality, we categorize features as dynamic or non-dynamic. Features describing standard deviation or max of values or features describing velocity are considered as dynamic features. Features describing averaged values or count numbers are non-dynamic. Thus for each unimodal, we have three feature sets in total. One includes all features in the modality, one includes all the dynamic features, and one includes all the non-dynamic features. In addition to unimodals, we also present multimodal feature sets that combine all three unimodals feature sets, all three dynamic feature sets, and all three non-dynamic feature sets.

6) Label: We use the mean of the observers rated *reaction* and *expressivity* as perceived expressivity score, then for each video, we would calculate the mean of all scores we received as ground truth. However, recall in Section III-B.2, the inter-rater reliability among the observers were acceptable but not great. To filter relatively less reliable observers, we computed modified z-scores using median absolute deviation (MAD) [45] of perceived expressivity scores for each video. Then for each crowd worker, we calculated the average of their 20 modified z-scores. Since our sample size was quite small (4-7 ratings for each video) when calculating the modified z-score, the common cutoff criteria of 3 MADs [46] would not be appropriate in our case. Instead of picking an arbitrary cutoff threshold, we removed 10% (N=28) observers with the highest modified z-scores on average (the least reliable observers in our task). This filtering procedure improved the ICC of perceived expressivity score to 0.87 with a 95% CI from 0.86 to 0.88 and ICC of overall items to 0.82 with a 95% CI from 0.81 to 0.82.

C. Methods

Since our expressivity labels were computed from a 7-point Likert scale, we formulated the task to predict perceived expressivity score as a regression problem. We used R^2 as our performance metric, which measures the proportion of the variance for a dependent variable that's explained by variables in a regression model. Finally, we experimented with two interpretable models.

1) Lasso: A shrinkage method for linear regression that penalizes the size of regression coefficients with L_1 regularization term. The amount of shrinkage is controlled by a constant factor λ [47]. Lasso is commonly used as a feature selection method. In a similar context [6], lasso was used to select features before a classifier was fit to predict a player's decision in IPD from game behaviors and facial expressions. Lasso is also easy to tune with one hyperparameter λ. We performed a grid search between 0 and 1 with a step of 0.1 to find the best λ.

2) Random Forest: Comparing to Lasso, random forests are an ensemble method for decision tree, which is equally interpretable without assuming a linear relationship between the features and the response. We build random forests by building several decision trees on bootstrapped training examples to reduce variance, thus avoid overfitting [48]. Two hyperparameters were tuned with grid search in our

experiments: number of trees (10, 30, 50) and max depth of the trees (4, 6, 8).

D. Experiments

For each of our feature sets, we used nested 10-fold cross-validation to train, validate, and evaluate our models [49]. In the inner loop, we performed grid search 10-fold cross-validation to tune hyperparameters and recorded R^2 for each fold. The best model was selected based on averaged R^2 during this validation process. We then test on the outer loop with another 10-fold cross-validation. In the outer loop, R^2 was recorded for each fold, and the 10 scores were averaged as our test score.

E. Results

Model Performance (R^2) are reported in Table I. First, we observe both models are suitable for our task, and random forests outperform lasso in all tasks. Second, dynamic models achieved comparable results to models with combined features in all modalities, and *multimodal* and *face* models outperformed *baseline* model. We can also see dynamic features alone are sufficient, and non-dynamic features do not contribute additional information to explain variance in perceived expressivity. Another important observation is that the face feature sets perform very close to the multimodal sets. Features from other modalities provide very little additional information. We examine the top 10 most important features for random forests with multimodal feature sets; the most important 8 features are related to the smile dynamics. The standard deviation (weight=0.21) and max of velocity (.14) of enjoyment smile factor (F1) are the most important two features. One posture dynamic feature (max velocity of overall landmarks, .02) and one head movement feature (standard deviation of overall head movement, .02) are also helpful, but the weights are relatively small.

TABLE I

	Lasso	Random Forests
Baseline	0.39	0.55
Multimodal	**0.63**	**0.67**
Multimodal Dynamic	**0.63**	**0.66**
Multimodal Non-dynamic	0.50	0.58
Face	**0.60**	**0.64**
Face Dynamic	**0.60**	**0.63**
Face Non-dynamic	0.39	0.55
Head	**0.30**	**0.36**
Head Dynamic	0.28	0.34
Head Non-dynamic	0.14	0.16
Posture	**0.16**	**0.28**
Posture Dynamic	**0.16**	**0.28**
Posture Non-dynamic	0.15	0.15

V. Application of Predicted Expressivity

Finally, we examined if our best-learned model of expressivity has value in recognizing important aspects of a social situation. First, we revisit a previously published finding which showed that smile intensity was associated with the unexpectedness of outcomes in social dilemmas [50]. Instead, we show that expressivity is a better predictor (and explanation) of this finding. Second, we provide preliminary evidence that expressivity is useful for identifying "moments of interest" in a video sequence.

A. Predicting Unexpectedness of Events

In recently published work, Lei and Gratch [50] made the claim that smile intensity was a good predictor of the unexpectedness of a decision in the iterated prisoner's dilemma (on the same corpus we use in this article)[3]. Specifically, they calculated the observed probability that a particular sequence of decisions occurred (for example, mutual splits were extremely likely if players mutually-split on the previous round, whereas a split-steal decision following a mutual-split was very unlikely). They found that the players' "enjoyment smile" intensity (as measured by the F1 feature described above), correlated highly with the unexpectedness of an event. Their analysis also showed that smiles were not correlated with whether the event was good or bad for the player, thus undermining the "standard model."

Here we examined if *expressivity* might better explain this finding. We reproduced the analysis following [50] with predicted expressivity from random forests models. Using predicted expressivity, we could improve the correlation (Pearson's r) with the unexpectedness of events (Table II).

TABLE II

	Smiles [50]	Multimodal	Face	Head	Posture
r	0.53*	0.76***	0.77***	0.67**	0.48

We see that the predicted expressivity from multimodal, face features, and head features are all better predictors than smiles. Recall in Fig. 2, when making judgments on expressivity, the extent to which the observers gained insights from the *head* correlates with insights gained from the *face* (r=0.69), indicating that facial movements are frequently co-occurring with head movements. Even though the head model was not as good as multimodal and face models in predicting the observers' perception, and head features contributed marginally in the multimodal model, the head model predicted perceived expressivity still conveys important information regarding the unexpectedness of the events players experienced.

B. Automatically Locating Interesting Segments in Videos

We perform a proof-of-concept exploration to examine if predicted expressivity could serve as a measure to automatically locate "moments of interest" within the entire video of an IPD game. To do this, we use our learned model of expressivity (the random forests model with all features) and calculate a moment-to-moment level of expressivity (moving a 3-second long sliding window, in 1-second increments, across the video). We then compare this continuous measure to known events within the game. Fig. 4 shows two examples.

[3]"surprise" was used in the original paper to represent the unexpectedness of events, to differentiate felt surprise from objectively measured event likelihood we think it is more accurate to use the term "unexpectedness of events" here

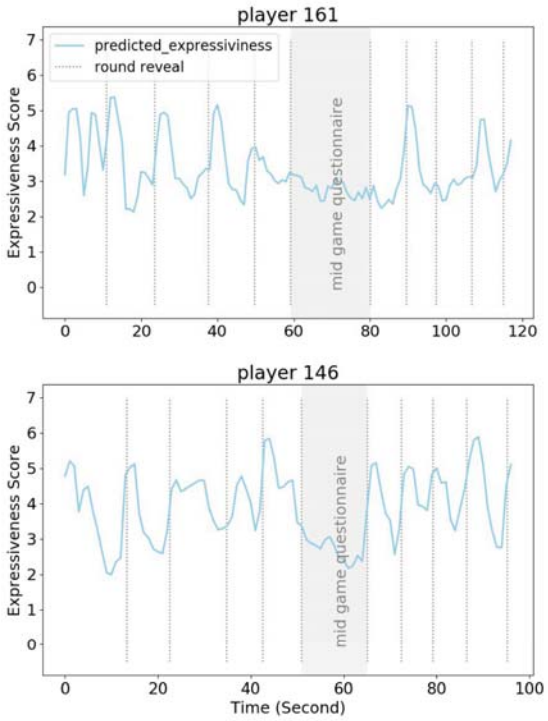

Fig. 4. Predicted expressivity peaks after the reveal moments.

Recall in Section III, the players played 10 rounds in each IPD game. In Fig. 4, each vertical line represents the moment that a joint decision was revealed to both player, following each round. The shaded area represents a moment in time when the game was interrupted so that players could complete a brief questionnaire about their subjective experience. We can see that players are most expressive at the moments when the joint-outcome is revealed. They are especially non-expressive when completing the questionnaire. Of course, this requires more systematic analysis but the results suggest the potential of the technique to automatically identify moments of interest.

VI. Discussion and Limitations

In this article, we examined if nonverbal expressivity could serve as a more useful construct for identifying significant moments within a social interaction, in contrast to assessing the presence or intensity of facial expressions. We collected observers' judgments on people's spontaneous reactions to significant events in an iterated prisoner's dilemma, built a model to predict observers' perception of expressivity, and with which to make inferences about the eliciting situation.

In terms of the research questions outlined at the start of this article, we can draw several conclusions. Concerning research questions (1) and (5), we provided evidence that expressivity may be a better indicator that something personally significant has happened to an individual, compared with using the intensity of individual facial actions. Specifically, we revisited a previously published claim that smile intensity was associated with the unexpectedness of outcomes in the IPD. Instead, we show that expressivity is a better predictor (and explanation) of this finding. (2) We showed that perceivers use more than facial expressions when making judgments of expressivity. Although, by constructing learned models, (3) we could show that perceivers mostly relied on dynamic features of facial expressions when making these judgments. (4) Finally, we were able to show that perceivers see these expressions as conveying thoughts as well as emotions, although perceivers mostly viewed these reactions as reflecting the senders' internal emotional state. From an algorithmic perspective, we next show that expressivity can be predicted with high accuracy, at least within the context of the IPD corpus.

There are several limitations to this current work. This research was only applied to a single corpus, and it remains to show that these expressivity findings generalize to other contexts and other corpora. An obvious next step would be to examine if the predictive models learned here can transfer to other spontaneous expression elicitation datasets. Our conclusions about the relative importance of different modalities must be qualified. Our model relies heavily on the accuracy of automatic facial expression estimation, face and head movement tracking. We used state-of-the-art trackers in this analysis, though they are not perfect. In particular, our approach to measuring posture is less validated when compared with our measures of facial actions. We are also interested in examining how end-to-end deep learning models would perform compared to the two simple interpretable models we presented. Though end-to-end models might be less interpretable, they can learn the best features given the task and do not require you to explicitly specify how different features are combined.

VII. Acknowledgments

This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

[1] A. Moors, "Flavors of appraisal theories of emotion," *Emotion Review*, vol. 6, no. 4, pp. 303–307, 2014.

[2] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[3] C. M. de Melo, P. J. Carnevale, S. J. Read, and J. Gratch, "Reading people's minds from emotion expressions in interdependent decision making." *Journal of personality and social psychology*, vol. 106, no. 1, p. 73, 2014.

[4] G. A. Van Kleef, "How emotions regulate social life: The emotions as social information (easi) model," *Current directions in psychological science*, vol. 18, no. 3, pp. 184–188, 2009.

[5] D. Vinkemeier, M. Valstar, and J. Gratch, "Predicting folds in poker using action unit detectors and decision trees," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 504–511.

[6] R. Hoegen, G. Stratou, and J. Gratch, "Incorporating emotion perception into opponent modeling for social dilemmas," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 801–809.

[7] P. Mussel, A. S. Göritz, and J. Hewig, "The value of a smile: Facial expression affects ultimatum-game responses." *Judgment & Decision Making*, vol. 8, no. 3, 2013.

[8] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, 2012.

[9] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[10] A. J. Fridlund, "The new ethology of human facial expressions," *The psychology of facial expression*, vol. 103, 1997.

[11] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[12] A. Scarantino, "How to do things with emotional expressions: The theory of affective pragmatics," *Psychological Inquiry*, vol. 28, no. 2-3, pp. 165–185, 2017.

[13] E. Krumhuber, A. S. Manstead, D. Cosker, D. Marshall, P. L. Rosin, and A. Kappas, "Facial dynamics as indicators of trustworthiness and cooperative behavior." *Emotion*, vol. 7, no. 4, p. 730, 2007.

[14] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological science*, vol. 16, no. 5, pp. 403–410, 2005.

[15] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological bulletin*, vol. 111, no. 2, p. 256, 1992.

[16] J. K. Burgoon and B. A. Le Poire, "Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality," *Communications Monographs*, vol. 66, no. 2, pp. 105–124, 1999.

[17] F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis." *Journal of Personality and Social Psychology*, vol. 71, no. 1, p. 110, 1996.

[18] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[19] R. T. Boone and R. Buck, "Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 163–182, 2003.

[20] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, no. 2, pp. 237–256, 2011.

[21] A. M. Kring and J. M. Neale, "Do schizophrenic patients show a disjunctive relationship among expressive, experiential, and psychophysiological components of emotion?" *Journal of abnormal psychology*, vol. 105, no. 2, p. 249, 1996.

[22] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and vision computing*, vol. 32, no. 10, pp. 641–647, 2014.

[23] D. A. Trevisan, M. Hoskyn, and E. Birmingham, "Facial expression production in autism: A meta-analysis," *Autism Research*, vol. 11, no. 12, pp. 1586–1601, 2018.

[24] A. L. Georgescu, B. Kuzmanovic, D. Roth, G. Bente, and K. Vogeley, "The use of virtual characters to assess and train non-verbal communication in high-functioning autism," *Frontiers in human neuroscience*, vol. 8, p. 807, 2014.

[25] R. Buck, C. Goldman, C. Easton, and N. Smith, "Social learning and emotional education," *Emotions in psychopathology: Theory and research*, pp. 298–314, 1998.

[26] L. Tickle-Degnen, "The interpersonal communication rating protocol: A manual for measuring individual expressive behavior," *Tufts University*, 2010.

[27] A. M. Kring, D. A. Smith, and J. M. Neale, "Individual differences in dispositional expressiveness: development and validation of the emotional expressivity scale." *Journal of personality and social psychology*, vol. 66, no. 5, p. 934, 1994.

[28] J. J. Gross and O. P. John, "Revealing feelings: facets of emotional expressivity in self-reports, peer ratings, and behavior." *Journal of personality and social psychology*, vol. 72, no. 2, p. 435, 1997.

[29] C. Neubauer, S. Mozgai, B. Chuang, J. Woolley, and S. Scherer, "Manual and automatic measures confirm—intranasal oxytocin increases facial expressivity," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 229–235.

[30] P. Wu, I. Gonzalez, G. Patsis, D. Jiang, H. Sahli, E. Kerckhofs, and M. Vandekerckhove, "Objectifying facial expressivity assessment of parkinson's patients: preliminary study," *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[31] A. Joshi, L. Tickle-Degnen, S. Gunnery, T. Ellis, and M. Betke, "Predicting active facial expressivity in people with parkinson's disease," in *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2016, p. 13.

[32] V. Lin, J. M. Girard, and L.-P. Morency, "Context-dependent models for predicting and characterizing facial expressiveness," *arXiv preprint arXiv:1912.04523*, 2019.

[33] R. H. Frank, *Passions within reason: The strategic role of the emotions.* WW Norton & Co, 1988.

[34] R. Hoegen, G. Stratou, G. M. Lucas, and J. Gratch, "Comparing behavior towards humans and virtual humans in a social dilemma," in *International Conference on Intelligent Virtual Agents*. Springer, 2015, pp. 452–460.

[35] T. Koo and M. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research. j chiropr med. 2016; 15 (2): 155–63."

[36] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242. Piscataway, NJ, 2003, pp. 133–142.

[37] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Face and gesture 2011*. IEEE, 2011, pp. 298–305.

[38] P. Ekman, "Facial action coding system," 1977.

[39] G. Stratou, J. Van Der Schalk, R. Hoegen, and J. Gratch, "Refactoring facial expressions: An automatic analysis of natural occurring facial expressions in iterative social dilemma," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 427–433.

[40] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[41] Z. Hammal, E. R. Wallace, M. L. Speltz, C. L. Heike, C. B. Birgfeld, and J. F. Cohn, "Dynamics of face and head movement in infants with and without craniofacial microsomia: An automatic approach," *Plastic and Reconstructive Surgery Global Open*, vol. 7, no. 1, 2019.

[42] S. Mohammad, K. Stefanov, S.-H. Kang, J. Ondras, and J. Gratch, "Multimodal Analysis and Estimation of Intimate Self-Disclosure," in *2019 International Conference on Multimodal Interaction*. New York, New York, USA: ACM Press, 2019, pp. 59–68.

[43] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[44] F. Trémeau, D. Malaspina, F. Duval, H. Corrêa, M. Hager-Budny, L. Coin-Bariou, J.-P. Macher, and J. M. Gorman, "Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects," *American Journal of Psychiatry*, vol. 162, no. 1, pp. 92–101, 2005.

[45] B. Iglewicz and D. C. Hoaglin, *How to detect and handle outliers*. Asq Press, 1993, vol. 16.

[46] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[48] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[49] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[50] S. Lei and J. Gratch, "Smiles signal surprise in a social dilemma," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 627–633.