Sayette Group Formation Task (GFT) Spontaneous Facial Expression Database

Jeffrey M. Girard[1], Wen-Sheng Chu[2], László A. Jeni[2], Jeffrey F. Cohn[1,2], Fernando De la Torre[2], and Michael A. Sayette[1]
[1] Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260
[2] Robotic Institute, Carnegie Mellon University, Pittsburgh, PA 15213

***Abstract*— Despite the important role that facial expressions play in interpersonal communication and our knowledge that interpersonal behavior is influenced by social context, no currently available facial expression database includes multiple interacting participants. The Sayette Group Formation Task (GFT) database addresses the need for well-annotated video of multiple participants during unscripted interactions. The database includes 172,800 video frames from 96 participants in 32 three-person groups. To aid in the development of automated facial expression analysis systems, GFT includes expert annotations of FACS occurrence and intensity, facial landmark tracking, and baseline results for linear SVM, deep learning, active patch learning, and personalized classification. Baseline performance is quantified and compared using identical partitioning and a variety of metrics (including means and confidence intervals). The highest performance scores were found for the deep learning and active patch learning methods. Learn more at http://osf.io/7wcyz.**

I. Introduction

Automated facial expression analysis is a growing area of research with numerous commercial and scientific applications ranging from consumer electronics and marketing to medicine and psychology. These applications capitalize on the central role that facial expressions have evolved to play in affective and interpersonal communication. It is no accident that the human brain devotes considerable resources to the analysis of faces [1], for the information they communicate is critical to many of life's most important endeavors.

The overwhelming majority of research on automated facial expression analysis uses one or another form of *supervised learning* [8], which requires the prior existence of annotated training data. For example, in order to train an algorithm to detect smiles in images, a large collection of images must be provided along with trusted labels (i.e., annotations) marking each image as a smile or non-smile.

The algorithm then attempts to learn a mapping between these images and the labels (e.g., smiles tend to have this pattern of characteristics, while non-smiles tend to have this other pattern of characteristics). This mapping then can be applied to automatically analyze novel images. Supervised learning can be quite successful, especially when the training images are sufficiently diverse [19] and when the algorithm is applied to novel images that are sufficiently similar to the ones it was trained on [20].

This research was supported in part by NIH grant MH096951. The Tesla K40 GPU used in this research was donated by the NVIDIA Corporation.

978-1-5090-4023-0/17/$31.00 ©2017 IEEE

However, the costs of collecting and annotating a large and diverse set of behavioral data are considerable. First, participants must be recruited, induced to produce facial expressions, recorded, and compensated. Then, facial expression annotators must be trained, supervised, and compensated for labeling the recordings. The process of labeling the images itself can be quite daunting. Annotating a single minute of video using the Facial Action Coding System (FACS) [14], the current gold-standard for labeling facial expressions [9], can require over an hour of annotators' time.

To circumvent these prohibitive costs, researchers have begun sharing their behavioral data and annotations in the form of facial expression databases. These databases are among the most impactful publications in the field, with many being cited hundreds or thousands of times. Early databases asked participants to pose different facial expressions in highly scripted and constrained settings. As facial expression analysis techniques advanced, these constraints were relaxed and more naturalistic databases were created.

At present, there are six publicly-available databases that contain recordings of spontaneous (i.e., non-posed) facial expressions and corresponding FACS annotations. Table I provides citations to these databases and details about their sizes and features. The median number of participants in these databases is 41 and the median number of video frames is 168,359. All six databases include FACS occurrence annotations (i.e., the binary presence or absence of facial actions), while only four include FACS intensity annotations (i.e., the ordinal magnitude of facial actions).

The most common context in which to record participants' facial behavior has been 'induced emotion,' which involves exposing participants to a variety of laboratory stimuli (e.g., sights, sounds, and smells) designed to elicit different types of emotion (e.g., amusement, surprise, or disgust). Other contexts include requiring participants to experience physical pain in a lab, to be interviewed by a virtual computer agent in a lab, and to watch television ads in their own homes.

Notably absent from this list of contexts is any form of social interaction between participants. This omission is likely a consequence of database creators' desire to present a standardized (i.e., controlled and consistent) context to all participants. It may also be that a solitary participant is the focal target of many applications of automated facial expression analysis (e.g., human-computer interaction).

However, humans are social creatures and many of the

TABLE I

PUBLICLY-AVAILABLE SPONTANEOUS FACIAL EXPRESSION DATABASES WITH FACS ANNOTATIONS

Dataset	Year	Participants	Frames	Occurrence	Intensity	Interaction	Context
UNBC-McMaster [26]	2011	25	48,398	•	•	–	Physical Pain
SEMAINE [30], [32]	2012	24	130,695	•	–	–	Artificial Listener
AM-FED [29]	2013	242	168,359	•	–	–	Market Research
DISFA [28]	2013	27	130,000	•	•	–	Induced Emotion
BP4D [35], [36]	2013	41	368,036	•	•	–	Induced Emotion
BP4D+ [37]	2016	140	1,400,000	•	•	–	Induced Emotion
GFT [Current]	2017	96	172,800	•	•	•	Group Formation

Note. Occurrence = FACS occurrence annotation; Intensity = FACS intensity annotation; Interaction = Multiple interacting participants.

important functions of facial expressions are related to social interaction [17]. As such, a database of multiple interacting participants is sorely needed to provide examples of how people actually behave during unscripted social interactions.

To address this gap in the literature, we present the Sayette Group Formation Task (GFT) database: a large and diverse database including 172,800 video frames from 96 participants, expert FACS annotations, meta-data, and baseline results. The novel context of the database is an unscripted social interaction within groups of three unacquainted adults.

II. PARTICIPANTS AND PROCEDURES

A. Recruitment and Demographics

Participants were drawn from a larger study on the impact of alcohol on group formation processes (for elaboration, see [31]). Healthy social drinkers between the ages of 21 and 28 were recruited via newspaper ads. To be included in the study, individuals had to affirm that they could comfortably drink at least three drinks in 30 min. Individuals were excluded if they reported a medical condition contraindicating alcohol consumption, met criteria for past alcohol abuse or dependence [3], were pregnant, or were more than 15 % above or below ideal weight for their height. This larger study included 720 participants (50 % female, 83 % white).

Participants in the current database represent a subset of the larger study's sample. These 96 participants (42 % female, 85 % white) were drawn from the set of participants whose audiovisual data were analyzable and who consented to having their audiovisual data used in further experiments. They were observed in groups of three participants, with mixed-gender groups ($n = 23$) being more common than same-gender groups ($n = 9$). Groups were randomly assigned to drink an alcoholic beverage ($n = 9$), a placebo beverage ($n = 8$), or a nonalcoholic control beverage ($n = 15$) during the experiment; all participants in a group drank the same type of beverage.

B. Experimental Setting

All participants were previously unacquainted. They first met after entering the observation room where they were seated around a circular table. They were asked to consume a beverage (based on their experimental condition) before engaging in a variety of cognitive tasks. We focus on a 1 min portion of the 36 min unstructured observation period during which participants became acquainted with one another; this portion started an average of 5.6 min into the observation period when participants were consuming the first of three equally-dosed drinks. They were asked not to discuss their level of intoxication, but could discuss any other topics.

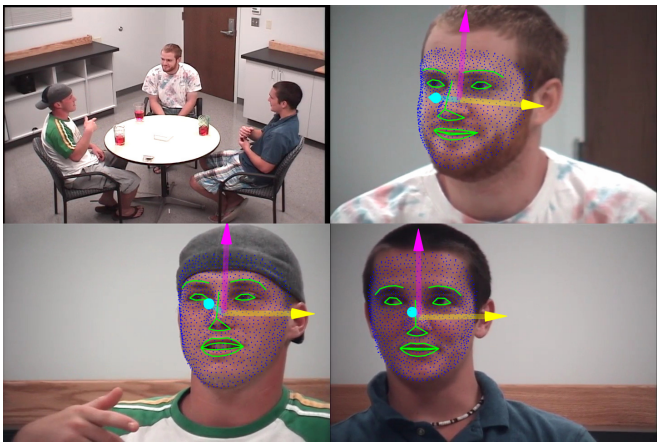

Fig. 1. Examples of video frames with facial landmarks and head pose

Separate wall-mounted cameras faced each participant. It was initially explained that the cameras were focused on their drinks and would be used to monitor participants' beverage consumption rates from the adjoining room; participants were later told of our interest in observing their behavior and all participants signed a second consent form indicating that they agreed to this use of their data.

III. DATA ACQUISITION AND ORGANIZATION

A. Recording Equipment

The observation room included a custom-designed video control system that permitted synchronized video capture for each participant, as well as an overhead shot of the group. Figure 1 provides an example frame from each camera. The video data collected by each camera had a standard frame rate of 29.97 frames per second and a resolution of 720×480 pixels. Audio was recorded from a single microphone.

B. Database Organization

The database is structured by group and by participant using unique identifiers for each group and participant. Each group is associated with video data from the group-level camera and with audio data recorded by the group-level microphone. Each participant is also associated with video

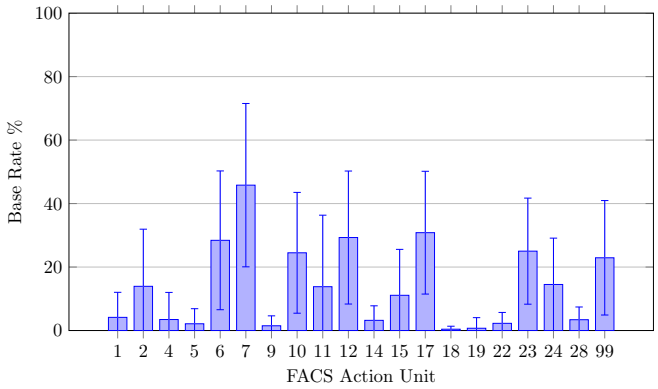

Fig. 2. Average FACS Base Rates with Standard Deviations

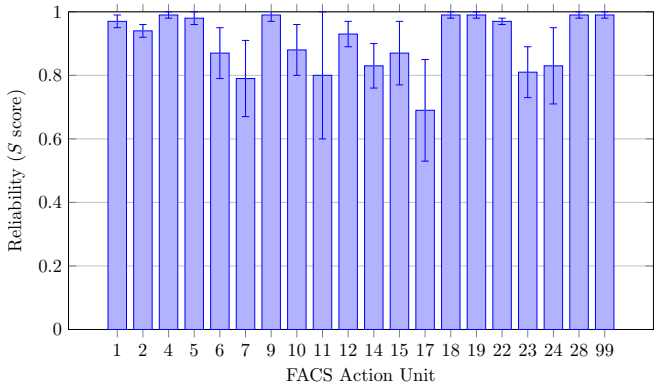

Fig. 3. Average Occurrence Reliability with 95% Confidence Intervals

data from a single wall-mounted camera. Synchronized video files from each camera are encoded as MPEG-4 files.

In addition, several types of annotation and meta-data are provided. Information about each participant's sex, ethnicity, and age is provided in a spreadsheet. Also included are the data-uses that each participant consented to (e.g., publishing images or showing video at conferences). Facial expression annotations, facial landmark tracking, and head pose estimation are provided for the participant-specific videos only.

With $1,800$ frames for each of 96 participants, the database contains a total of $172,800$ frames. The total file size of the database is 1.4 GB. The database will be made available to researchers from `http://osf.io/7wcyz`. The terms of use for the database require researchers to respect the data-uses that each participant consented to. Specifically, one participant did not want their images used in publications and three participants did not want their videos shown at scientific meetings.

IV. Data Annotation and Meta-Data

A. FACS Occurrence Annotation

The FACS manual [14] defines 32 distinct facial action units (AUs). Twenty AUs that commonly occurred in this dataset and that are implicated in affective and interpersonal communication were manually coded. Occurrence coding involves assigning each video frame to one of two categories:

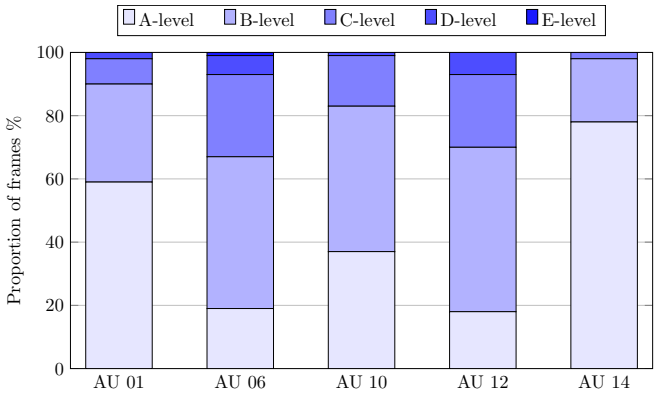

Fig. 4. Distribution of FACS Intensity Levels

present (i.e., contains a given AU) or *absent* (i.e., does not contain that AU). Frames were coded as present only if they contained the AU at the B-level of intensity or higher (see subsection IV-B). AUs were annotated during speech.

The distribution of occurrence codes was evaluated using base rates (Figure 2). The base rate of an AU is equal to the number of present frames divided by the total number of valid (i.e., non-occluded) frames. While some AUs occurred more than $30\,\%$ of the time (AUs 6, 7, 10, 12, and 17), other occurred less than $5\,\%$ of the time (AUs 4, 5, 9, 18, 19, 22, and 28). Occlusions (AU 99), or frames in which the face was partially obstructed from view, occurred $22\,\%$ of the time and were not coded. As shown by the standard deviation error bars in this figure, the base rates for some AUs varied greatly between participants.

To assess inter-rater reliability, a subset of participants was selected at random to be annotated by multiple coders. By comparing annotations between coders on a frame-by-frame basis, we can quantify the degree to which they tend to be consistent in the assignment of frames to categories. Here we estimate reliability using the free marginal kappa coefficient (i.e., S score) [4], [5]. For more information on this index, see subsection V-C on baseline method evaluation.

Figure 3 depicts the S score, averaged across 23 participants, for each AU. The error bars depict $95\,\%$ confidence intervals. Note that random guessing (e.g., flipping a coin) would yield an S score of 0.00 and perfect agreement would yield a score of 1.00. Although such thresholds can be oversimplifying, a rule of thumb suggests that kappa-like scores between 0.60 and 0.80 are "good," while scores above 0.80 are "very good" [2]. Average occurrence reliability scores were all above 0.60. However, we note that the confidence interval for AU 17 extended below this threshold.

B. FACS Intensity Annotation

Five AUs were selected for intensity coding due to their importance in affective and interpersonal signaling, including AUs 1, 6, 10, 12, and 14. To accomplish intensity coding, new videos were created by concatenating video frames that had been annotated as present by occurrence coders (frames within a small range around present frames were also

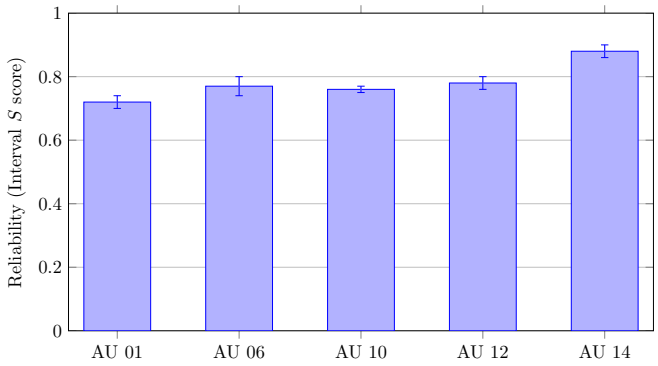

Fig. 5. Average Intensity Reliability with 95% Confidence Intervals

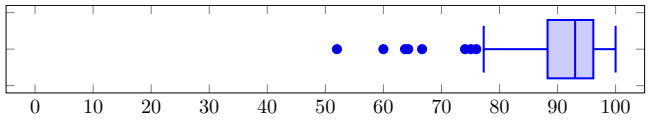

Fig. 6. Percentage of Frames with Tracking Results (with all outliers)

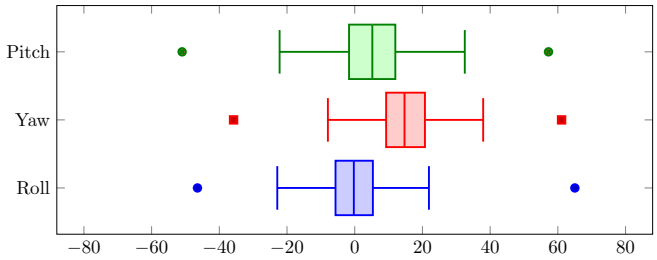

Fig. 7. Head Pose Distributions (in degrees, with extrema outliers)

included for context). Intensity coders then viewed these new videos and, using rules from the FACS manual, annotated each frame by assigning it to one of seven categories: absent, A-level, B-level, C-level, D-level, E-level, or uncodeable. Thus, it was possible for the intensity and occurrence coders to disagree on a particular frame.

The distribution of intensity levels was similar across AUs. Overall, A-level (42 %) and B-level (39 %) frames were most common, with C-level (15 %) frames being rare and and D-level (3 %) and E-level (1 %) frames almost never occurring. The proportion of A-level frames was particularly high for AUs 1 and 14, while the proportion of C-level and D-level frames was particularly high for AUs 6 and 12.

Assessing inter-rater reliability for intensity coding is similar to assessing inter-rater reliability for occurrence coding. However, a reliability index for intensity codes needs to account for the fact that some categories are more similar than others (e.g., A-level and B-level are more similar than B-level and E-level are). Here we estimate the reliability of intensity coding using a generalization of the S score that allows for different "weights" or degrees of similarity to be provided for each pair of categories [21]. We use "linear/interval" weights, which assume equal spacing between categories and are more conservative than "ordinal" weights which do not.

Between 29 and 79 participants (per AU) were randomly selected to be intensity coded by multiple coders. Reliability, as quantified by the interval-weighted S score, averaged 0.78. Reliability for individual AUs ranged from 0.72 to 0.88. The confidence interval error bars in Figure 5 show that our sample-based estimates of reliability are very likely accurate within an average of 0.02 points. That all the average S scores exceeded 0.60 suggests "good" reliability overall.

C. Facial Landmark Tracking and Head Pose Estimation

To track the location of important facial landmarks (e.g., eyes, brows, nose, and mouth) in the participant-specific videos, we used ZFace [23], a real-time face alignment software which accomplishes dense 3D registration from 2D videos and images without requiring person-specific training.

ZFace first estimates the location and visibility of a dense set of facial landmarks and then reconstructs face shapes by fitting a part-based 3D model. This model includes parameters for scale, translation, non-rigid transformation, and rotation in three dimensions (i.e., pitch, yaw, and roll). Figure 1 shows a visualization of the landmarks (as blue and green dots) and the rotation parameters (as 3D arrows). Figure 6 shows the percentage of tracked video frames, and Figure 7 depicts the estimated head pose distributions.

V. Baseline Methods

One of the main goals of this database is to provide a standardized set of behavioral data that researchers can use for comparing methods of automatic facial expression analysis. To encourage and facilitate such comparisons, we describe and present the results of "baseline" methods that researchers can compare their own results to. For AU occurrence detection, we include baseline results from a maximum-margin framework and a deep learning framework. Of the 20 AUs that were annotated, we selected the 10 that both occurred more than 5 % of the time and showed "very good" reliability among human coders with S scores greater than 0.80 (see Table II for the list of selected AUs). Baseline results for AU intensity estimation are being developed and will be added in a planned expansion of the current paper.

The entire data set was partitioned into three subsets: a training set composed of 20 groups (60 participants), a validation set composed of 6 groups (18 participants), and a testing set composed of 6 groups (18 participants). Groups were assigned to partitions in order to maximize the similarity between the partitions; as such, the AU distributions are very similar in each of the three sets (i.e., Bhattacharyya coefficients > 0.99). This partitioning scheme was used by both baseline methods and is included in the data distribution so that others can replicate it in their own experiments.

For both the maximum-margin and deep learning frameworks, initial training was completed using the training set and parameters were tuned using the validation set. Finally, performance scores were calculated using the testing set. Using independent partitions for each step helps maximize generalizability and helps prevent overfitting.

A. Maximum-Margin Framework

Maximum-margin classifiers attempt to learn a class-separating hyperplane that maximizes the distance or margin between the hyperplane and the nearest data points. This framework is well-accepted for AU detection due to its competitive performance and high efficiency. As a standard baseline from the maximum-margin framework, we include a linear support vector machine (SVM) approach. We also include experiments using extensions of this approach, including active patch learning and personalized classification. All experiments were evaluated using the same set of features and the same data partitions. Below, we describe pre-processing and each approach in turn.

Pre-processing: Using the tracked facial landmarks and a similarity transformation, each video frame was registered to a canonical view with the size of the face normalized to have an inter-ocular distance of 100 pixels. HOG descriptors [11] were then extracted around each of the 49 landmarks using 64×64 pixel patches divided into 16 cells and 8 orientation bins. The feature vector for each image had a total of $6,272$ elements (i.e., $49 \times 16 \times 8$); features were normalized to have zero mean and unit variance. To create a balanced distribution of training examples, participant-based undersampling was used, i.e., all examples of the minority class were included from each participant and then an equally sized subsample of the majority class was added.

Linear SVM: After the HOG features were extracted, a two-class linear SVM was trained for each AU using the LIBLINEAR open-source library [15]. The regularization parameter, C, was tuned within $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$.

Active patch learning (APL): Observe that the features are structured and locally dependent, i.e., every 128 values of the feature vector can be treated as a "group" because these values were extracted from the same facial patch. Given this fact, as well as the knowledge that AUs correspond to motion on specific regions of the face, the APL approach [40], [38] aims to automatically select a sparse subset of facial patches for recognizing each AU (e.g., only selecting patches around the eyebrows when classifying the presence or absence of AU 1, which raises the inner portion of the brow). Specifically, we implemented APL as a logistic regression regularized by ℓ_2 group lasso. We used the SPAMS toolbox [27] and tuned the λ parameter within $\{2^{-10}, 2^{-9}, \ldots, 2^{-1}\}$.

Personalized classification: Individual differences in facial morphology and physiology can substantially influence the performance of AU detection algorithms by creating overlap between classes in high-dimensional feature space. For example, a person-independent classifier may mistakenly classify an individual with naturally upturned eyebrows as constantly showing AU 1, which produces this effect in individuals with naturally flat eyebrows. Accounting for such differences has become a large area of research interest (e.g., [34]). To test the influence of such factors in the GFT database, we performed experiments using the Selective Transfer Machine (STM) approach [6]. The C and λ parameters were each tuned within $\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$.

B. Deep Learning Framework

Deep learning is a biologically inspired approach that attempts to mimic the activity of layers of neurons in the brain. In recent years, this approach has produced dominating performance in many learning tasks (e.g., object recognition). Compared to the maximum-margin framework, deep learning offers a number of benefits. First, the highly nonlinear nature of its network infrastructure enables it to capture the richness and diversity of complex data. Second, its stochastic training procedure allows it to include and learn from truly large amounts of training data. Finally, it replaces handcrafted features with algorithms for data-driven feature learning.

As a baseline, we adopted AlexNet [25] by modifying its output layer to accommodate multi-label output. Given an expert-annotated label $\mathbf{y} \in \{-1, 0, +1\}^L$ for L AUs (-1/$+1$ indicates absence/presence of an AU, and 0 missing label) and a prediction $\widehat{\mathbf{y}} \in \mathbb{R}^L$, this multi-label network aims to minimize the multi-label cross entropy loss:

$$L(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{-1}{L} \sum_{\ell=1}^{L} [y_\ell > 0] \log \widehat{y}_\ell + [y_\ell < 0] \log(1 - \widehat{y}_\ell),$$

where $[x]$ is an indicator function returning 1 if x is true, and 0 otherwise. The proposed multi-label architecture is similar to [18], which takes 40×40 pixel images as input. However, we used 200×200 pixel images in order capture more detail regarding facial texture that may aid in recognizing AUs.

We trained the multi-label network with batches of 512 samples, 30 epochs, a momentum of 0.9, and a weight decay of 0.01. All models were initialized with a learning rate of 0.001, which was further reduced manually after five training epochs. The implementation was based on the Caffe toolbox [24] with modifications to support the multi-label cross entropy loss. To obtain the predicted labels $\widehat{\mathbf{y}}$, the sign function was used as an activation function.

C. Performance Evaluation

Given that different performance metrics are preferred by different researchers and often focus on different aspects of the task, several metrics are included for completeness. For all metrics, we include an estimate of performance (i.e., the mean of performance scores for each participant in the testing set) as well as a 95% confidence interval to represent the precision of that estimate; a confidence interval can be considered a range of highly plausible values [10].

First, the S score or "free-marginal kappa coefficient" is included as an overall, chance-adjusted summary statistic [4], [5]. It estimates chance agreement by assuming that each category is equally likely to be chosen at random [39]. When applied to two raters assigning objects to dichotomous categories, the S score is calculated using (1), where n_{kl} is the number of objects that the first rater assigned to category k and that the second rater assigned to category l.

$$\mathrm{S} = \frac{2n_{00} + 2n_{11}}{n_{00} + n_{11} + n_{01} + n_{10}} - 1 \qquad (1)$$

TABLE II

AU-SPECIFIC BASELINE RESULTS FOR AU OCCURRENCE DETECTION WITH 95% CONFIDENCE INTERVALS

	Maximum-Margin Framework (Linear SVM)				Deep Learning Framework			
AU	S	PA	NA	AUC	S	PA	NA	AUC
01	$.79 \pm .17$	$.38 \pm .11$	$.92 \pm .09$	$.88 \pm .10$	$.86 \pm .09$	$.44 \pm .19$	$.95 \pm .04$	$.88 \pm .09$
02	$.61 \pm .15$	$.32 \pm .14$	$.81 \pm .13$	$.80 \pm .05$	$.59 \pm .25$	$.46 \pm .18$	$.82 \pm .14$	$.81 \pm .09$
04	$.56 \pm .29$	$.13 \pm .13$	$.85 \pm .13$	$.72 \pm .16$	$.85 \pm .10$	$.02 \pm .04$	$.96 \pm .03$	$.60 \pm .14$
06	$.69 \pm .11$	$.67 \pm .14$	$.85 \pm .10$	$.91 \pm .06$	$.72 \pm .14$	$.73 \pm .11$	$.89 \pm .06$	$.93 \pm .05$
10	$.72 \pm .12$	$.64 \pm .13$	$.89 \pm .05$	$.92 \pm .05$	$.79 \pm .09$	$.72 \pm .10$	$.92 \pm .04$	$.94 \pm .04$
12	$.71 \pm .11$	$.78 \pm .10$	$.86 \pm .07$	$.94 \pm .04$	$.74 \pm .15$	$.82 \pm .07$	$.87 \pm .08$	$.97 \pm .02$
14	$.70 \pm .13$	$.15 \pm .07$	$.91 \pm .05$	$.77 \pm .11$	$.93 \pm .05$	$.05 \pm .09$	$.98 \pm .01$	$.79 \pm .12$
15	$.58 \pm .15$	$.29 \pm .11$	$.86 \pm .07$	$.75 \pm .08$	$.80 \pm .12$	$.19 \pm .11$	$.94 \pm .05$	$.77 \pm .07$
23	$.27 \pm .18$	$.49 \pm .09$	$.67 \pm .11$	$.73 \pm .06$	$.56 \pm .12$	$.43 \pm .11$	$.85 \pm .05$	$.77 \pm .07$
24	$.52 \pm .15$	$.44 \pm .10$	$.83 \pm .06$	$.86 \pm .06$	$.77 \pm .08$	$.42 \pm .13$	$.93 \pm .03$	$.85 \pm .06$

Note. S = free-marginal kappa, PA = positive agreement (equal to F_1 here), NA = negative agreement, AUC = area under ROC.

Next, positive agreement (PA) and negative agreement (NA) are included as category-specific performance measures [7], [12]. Collectively, these metrics are referred to as "specific agreement." When applied to two raters, the interpretation of specific agreement is the probability of one rater assigning an object to a specific category given that the other rater has also assigned the object to that category. In the case of two raters and dichotomous categories, PA is equal to the F_1 score. We refer to this metric as PA rather than F_1, despite the popularity of the latter name, because PA (or specific agreement) is a more generalized metric that can be applied in situations with any number of raters and categories. These scores can be calculated using (2) and (3).

$$\text{PA} = \frac{2n_{11}}{2n_{11} + n_{10} + n_{01}} \tag{2}$$

$$\text{NA} = \frac{2n_{00}}{2n_{00} + n_{10} + n_{01}} \tag{3}$$

Finally, AUC or area under the receiver-operating characteristic (ROC) curve is included as a threshold-independent measure of performance [16]. It is equal to the probability that the classifier will rank a randomly chosen positive object higher than a randomly chosen negative object. As seen in (4), it is equal to the integral of the product of each threshold's (T) true positive rate and false positive rate. It can be estimated using multiple trapezoidal approximations.

$$\text{AUC} = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT \tag{4}$$

Visualizations of classifier performance are also provided in the form of cost curves [13] (Figure 8). Cost curves are related to ROC curves in that each point in ROC space is represented by a line in cost space. However, cost curves allow easier visual comparison between classifiers, enable the calculation of confidence intervals, and provide insights on a classifier's performance over varying class probabilities and misclassification costs [22]. The closer a cost curve is to the bottom of the graph, the better performance is; "trivial classifiers" are represented as diagonals in cost space, showing the performance of assigning all objects to the positive category or the negative category, respectively.

TABLE III

MEAN BASELINE RESULTS WITH 95% CONFIDENCE INTERVALS

Method	S	PA	NA	AUC
SVM	$.61 \pm .07$	$.45 \pm .06$	$.84 \pm .04$	$.83 \pm .03$
APL	$.74 \pm .05$	$.40 \pm .07$	$.90 \pm .03$	$.83 \pm .02$
STM	$.62 \pm .05$	$.45 \pm .05$	$.85 \pm .04$	$.82 \pm .03$
DL	$.76 \pm .05$	$.46 \pm .07$	$.91 \pm .03$	$.84 \pm .03$

Note. Means were calculated by averaging within then across participants.

VI. BASELINE RESULTS AND DISCUSSION

The results of our baseline experiments are presented in several tables and figures. Table II provides AU-specific results for our two primary baselines: linear SVM and deep learning; Figure 8 depicts these results as cost curves. Table III provides summary results (i.e., means with 95% confidence intervals) for all four approaches. Due to space constraints, the AU-specific results for the APL and STM are provided in the supplementary material.

We begin our discussion of the baseline results by comparing the linear SVM and deep learning approaches. While these approaches were not significantly different in terms of mean PA ($\Delta = .00$, 95 % CI: $[-.07, .07]$) or mean AUC ($\Delta = .02$, $[-.02, .05]$), deep learning had a significantly higher mean S ($\Delta = .15$, $[.09, .21]$) and a significantly higher mean NA ($\Delta = .07$, $[.04, .10]$) than linear SVM. Thus, deep learning had better "overall" performance and this was largely driven by increased accuracy on the negative class. This increased ability to tell when AUs were absent may be due to deep learning's feature-learning capabilities or its multi-label loss function (i.e., ability to model the dependencies between AUs).

Visual inspection of Table II and Figure 8 reveals that performance varied greatly between AUs. The AUs that were most successfully detected were 6, 10, and 12; these AUs showed the lowest cost curves, the highest AUC scores, and uniquely high PA scores. These AUs had relatively high base rates (i.e., 20 % to 40 %), but this does not appear to be sufficient for good performance as AU 23 was also quite common and yet proved difficult to detect for both methods. Other AUs, such as 4 and 14, were less common

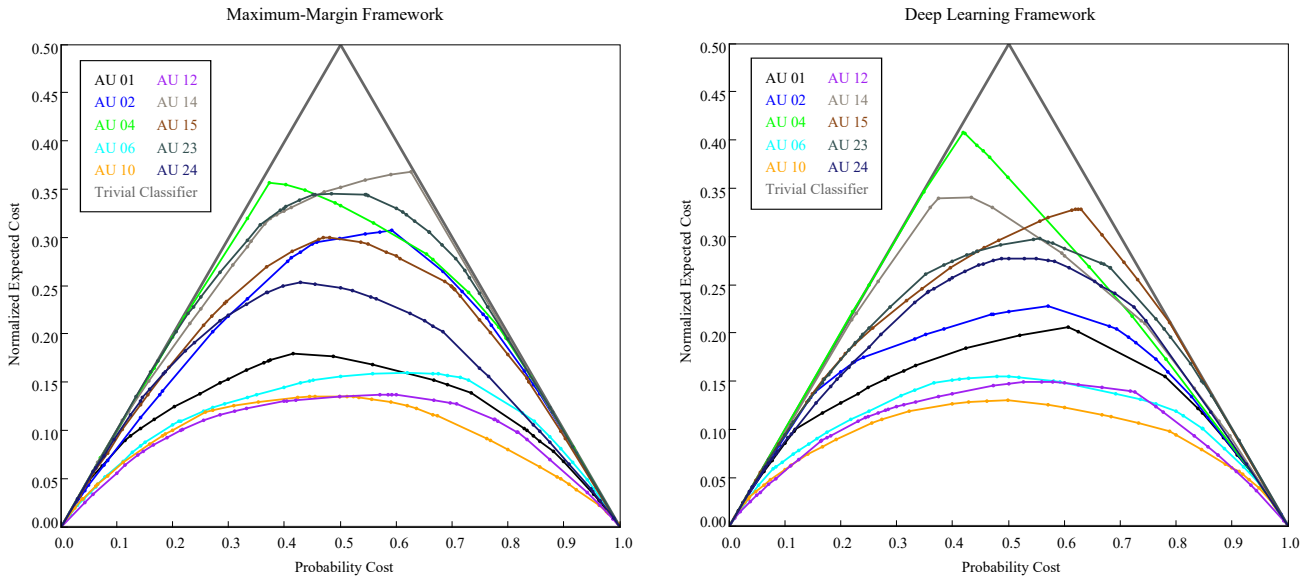

Fig. 8. AU-Specific Cost Curves for Baseline Results

and also very difficult to detect. The cost curves for AU 4, for example, show that performance was virtually no better than that of a trivial classifier until a probability cost of 0.4. This means that, unless applied in a setting where AU 4 occurs 40 % of the time or more, you might as well use a classifier that always predicts that the AU is absent (cf. the observed base rate for AU 4 in this sample was 3.5 %).

In addition to our two main baselines, we also ran several experiments to determine the effectiveness of expansions to the linear SVM method. Given the high efficiency of the maximum-margin framework, an expansion capable of meeting or exceeding the performance of deep learning would be highly desirable. We first tested APL, which accounts for spatial relationships among features that linear SVM ignores. We then tested STM, which accounts for individual differences in facial morphology and behavior.

APL was not significantly different from linear SVM in terms of mean AUC ($\Delta = .00$, $[-.02, .02]$). However, it did have a significantly higher mean S ($\Delta = .13$, $[.08, .17]$), a significantly higher mean NA ($\Delta = .06$, $[.04, .08]$), and a significantly *lower* mean PA ($\Delta = -.05$, $[-.09, -.01]$) than linear SVM. APL was also not significantly different from deep learning according to any metric: mean S ($\Delta = -.02$, $[-.07, .02]$), PA ($\Delta = -.06$, $[-.13, .02]$), NA ($\Delta = -.01$, $[-.03, .01]$), or AUC ($\Delta = -.01$, $[-.05, .02]$). Thus, APL appears to have traded away some of its PA in exchange for NA. Because the negative class is generally more common, this trade also led to a higher overall S score. However, it is difficult to say for sure that this is a worthwhile trade. Ideally, PA and NA would both increase using the same method.

STM was not significantly different from linear SVM according to any metric: mean S ($\Delta = .01$, $[-.05, .07]$), PA ($\Delta = .00$, $[-.03, .03]$), NA ($\Delta = .00$, $[-.02, .03]$), or AUC ($\Delta = -.01$, $[-.03, .01]$). This is a surprising result given that STM has outperformed linear SVM previously [6]. Two differences in the current work may account for this disparity. First, the current work used feature vectors with far higher dimensionality than did previous work (i.e., thousands versus hundreds of features), which may have produced greater class separability for the linear SVM to capitalize on. Second, the current work included a larger amount of training data than did previous work, which may have allowed the linear SVM to avoid overfitting. Thus, STM may be better suited to use when features and amounts of training data are insufficient.

VII. General Discussion

Publicly-available facial expression databases have been integral to the development and refinement of approaches for the automatic analysis of facial behavior. Recent databases have prioritized the collection of spontaneous (non-posed) expressions and expert annotations of AU intensity. However, despite the role that facial behavior plays in social communication, no such databases have included multiple interacting participants. The current paper describes the Sayette Group Formation Task (GFT) database, which addresses this gap in the literature by providing examples of how participants behave and communicate with one another during an unscripted, small-group interaction. Relative to previous databases, GFT includes a high number of participants and a comparable number of video frames. It also includes expert annotations of FACS occurrence and intensity, facial landmark tracking, and numerous baseline results.

This large and diverse dataset also provides an excellent opportunity to evaluate the generalizability and scalability of different methods for AU detection. Our experiments point toward three considerations that researchers would be wise to keep in mind while designing AU detection algorithms. First, facial structure and musculature introduce spatial dependencies within feature space that can influence learning. Second, participants' individual differences in facial morphology and

physiology can influence learning. Lastly, interactions and correlations between AUs can influence learning.

Several characteristics of the database warrant additional discussion. First, the age range of participants was restricted to $21-28$ years old, which may limit generalizability beyond young adults. Second, the majority of participants ($85\,\%$) were white, which may limit generalizability to non-white participants. Lastly, the majority of AU occurrence frames had low intensity levels. We believe this to be a context effect, i.e., that spontaneously produced expressions tend to be quite subtle during unscripted social interactions.

Several directions for future work on this database are planned. First, we plan to expand the current paper by adding baseline results for the automatic estimation of AU intensity. Second, we plan to add new types of expert annotation, such as continuous and dimensional ratings of affect and interpersonal behavior. Third, we have reserved a hold-out set of 54 participants, which could be contributed to a public challenge in the future (e.g., FERA [32] or AVEC [33]).

REFERENCES

[1] R. Adolphs. The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 60:693–716, 2009.

[2] D. G. Altman. *Practical statistics for medical research*. Chapman and Hall, 1991.

[3] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Author, Washington, DC, 4th edition, 1994.

[4] E. M. Bennett, R. Alpert, and A. C. Goldstein. Communication through limited response questioning. *The Public Opinion Quarterly*, 18(3):303–308, 1954.

[5] R. L. Brennan and D. J. Prediger. Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3):687–699, 1981.

[6] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[7] D. V. Cicchetti and A. R. Feinstein. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558, 1990.

[8] J. F. Cohn and F. De la Torre. Automated face analysis for affective computing. In R. A. Calvo, S. K. D'Mello, J. Gratch, and A. Kappas, editors, *Handbook of affective computing*. Oxford, New York, NY, 2014.

[9] J. F. Cohn and P. Ekman. Measuring facial action. In J. A. Harrigan, R. Rosenthal, and K. R. Scherer, editors, *The new handbook of nonverbal behavior research*, pages 9–64. Oxford University Press, New York, NY, 2005.

[10] G. Cumming and S. Finch. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61(4):532–574, 2001.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[12] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[13] C. Drummond and R. C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

[14] P. Ekman, W. V. Friesen, and J. Hager. *Facial action coding system: A technique for the measurement of facial movement*. Research Nexus, Salt Lake City, UT, 2002.

[15] R.-E. Fan, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[16] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[17] A. J. Fridlund. *Human facial expression: An evolutionary view*. Academic Press, 1994.

[18] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *ACII*, 2015.

[19] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre. How much training data for facial action unit detection? In *FG*, 2015.

[20] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre. Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behavior Research Methods*, 47(4):1136–1147, 2015.

[21] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, Gaithersburg, MD, 4th edition, 2014.

[22] H. He and E. A. Garcia. Learning from imbalanced data sets. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1264, 2010.

[23] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d video for real-time use. *Image and Vision Computing*, 2016.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint, arXiv:1408.5093*, 2014.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[26] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. *FG*, pages 57–64, 2011.

[27] J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[28] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[29] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. *CVPR*, pages 881–888, 2013.

[30] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[31] M. A. Sayette, K. G. Creswell, J. D. Dimoff, C. E. Fairbairn, J. F. Cohn, B. W. Heckman, T. R. Kirchner, J. M. Levine, and R. L. Moreland. Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding. *Psychological Science*, 23(8):869–878, 2012.

[32] M. F. Valstar, T. Almaev, J. M. Girard, G. Mckeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In *FG*, 2015.

[33] M. F. Valstar, B. W. Schuller, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014: The 4th international audio/visual emotion challenge and workshop. *ACM Multimedia*, pages 1243–1244, 2014.

[34] J. Zeng, W.-S. Chu, F. D. la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. *IEEE Transactions on Image Processing*, 25(10):4753–4767, 2016.

[35] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3D dynamic facial expression database. *FG*, 2013.

[36] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[37] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, pages 3438–3446, 2016.

[38] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946, 2016.

[39] X. Zhao, J. S. Liu, and K. Deng. Assumptions behind inter-coder reliability indices. In C. T. Salmon, editor, *Communication Yearbook*, pages 418–480. Routledge, 2012.

[40] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012.