Context-sensitive Prediction of Facial Expressivity using Multimodal Hierarchical Bayesian Neural Networks

Anonymous FG 2018 submission
Paper ID ****

***Abstract*— Objective automated affect analysis systems can be applied to quantify the progression of symptoms in neurodegenerative diseases such as Parkinson's Disease (PD). PD hampers the ability of patients to emote by decreasing the mobility of their facial musculature, a phenomenon known as "facial masking." In this work, we focus on building a system that can predict an accurate score of active facial expressivity in people suffering from Parkinson's disease using features extracted from both video and audio. An ideal automated system should be able to mimic the ability of human experts to take into account contextual information while making these predictions. For example, patients exhibit different emotions with varying intensities when speaking about positive and negative experiences. We utilize a hierarchical Bayesian neural network framework to enable the learning of model parameters that subtly adapt to pre-defined notions of context, such as the gender of the patient or the valence of the expressed sentiment. We evaluate our formulation on a dataset of 772 20-second video clips of Parkinson's disease patients and demonstrate that training a context-specific hierarchical Bayesian framework yields an improvement in model performance in both multiclass classification and regression settings compared to the same model trained on all data pooled together.**

I. INTRODUCTION

Automatic and accurate affect sensing can play a major role in diagnostic as well as treatment procedures in medical conditions where emotive, expressive and cognitive abilities are impaired. The development of such technologies can aid therapists and practitioners, for example, by helping them save valuable time otherwise devoted to laborious manual coding of patient observations. The impressive progress made in the field of automatic facial expression analysis [8], [14], [30] has spurred computational research in applications related to healthcare and behavioral psychology. In this work, we focus on developing a machine learning model capable of predicting facial expressivity ratings of patients with Parkinson's disease (PD) from short interview videos.

PD affects over 10 million people worldwide and about 1% of people over 60 years old [1]. Patients with Parkinson's disease (PD) often have a reduced ability to exhibit spontaneous facial expression due to an increased rigidity of facial musculature, also known as facial masking [28] or facial bradykinesia [5]. The reduced ability in patients to express emotions can hinder aspects of their social life because they are often misperceived by others [28]. It is therefore important for clinicians and researchers to be able to objectively assess and quantify the level of active expressivity in the face, so they can measure facial masking as a symptom of PD and test whether interventions to improve facial masking are effective.

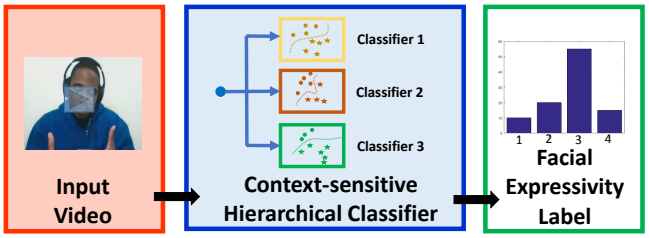

Fig. 1. Given a dataset consisting of short audio-video clips of PD patient interviews along with expertly annotated facial expressivity labels, relevant features are extracted from both video and audio modalities. A hierarchical model is trained that leverages additional contextual information to predict active facial expressivity ratings of the patients. We experiment with two notions of context: *gender*: a variable specifying the gender (male or female) of the patient and *sentiment*: a variable specifying the sentiment (positive or negative) expressed in the interview clip.

Facial expressivity is inherently more difficult to measure in people with PD because facial masking dims the clarity of muscle action shown in the face. Despite this difficulty there are existing manuals for objectively measuring active expressivity in the parkinsonian face, one of which is the Interpersonal Communication Rating Protocol (ICRP) [26], where active facial expressivity is among 20 indicators rated by trained experts along a 5-point Likert scale. Raters are trained to provide a "Gestalt" rating based on the intensity (strength of emotion or movement), duration (how long a behavior or movement lasts) and frequency (how often a behavior or movement lasts) of the expressive behavior observed. An active facial expressivity rating of 1 represents a person with "primarily one emotional expression plastered on the face, with low to no movement" whereas a rating of 5 is given to people with "highly active, animated, mobile and moving face with changing emotional expressions" [26].

As with other systems of manual coding, rating facial expressivity according to the ICRP brings forth challenges associated with scale and feasibility. Human coders have successfully coded facial expression in people with PD [16], but the costs associated with the manual assessment of all patients with PD can be prohibitively high. Comprehensive manual coding of 20 seconds of video can take upwards of an hour, and often two coders are needed to establish that the human coder is reliable.

Existing works involving computational analyses of facial emotions and expressivity of PD patients are mostly limited to pilot studies comparing facial characteristics and dynamics between a small group of PD patients and a separate control

group [1], [4], [33]. An accurate machine learning model trained on an expertly annotated dataset and capable of generalizing to new data is, therefore, an attractive proposition. In this work, we utilize a dataset of 772 short interview audio-video clips of PD patients and their corresponding facial expressivity labels to train a model that can accurately predict facial expressivity levels of new PD patients. For each video, we extract interpretable visual features from the face of the patients detected in the input frames as well as audio features from the raw audio to produce a multimodal feature descriptor, with which we train both classification and regression models.

Most existing works on automated facial analysis train generic classifiers for the task-at-hand, ignoring additional context that can accompany the input. Contextual information can be derived, for example, from the identity of the patient, the gender of the patient, the mood of the patient during the time of the interview, etc. In this work, we investigate whether contextual information can be leveraged to further improve the performance of the model. We experiment with two clearly defined notions of context: (1) *gender*: a variable indicating the gender (male or female) of the patient and, (2) *sentiment*: a variable indicating the sentiment (positive or negative) expressed during the interview. These variables are provided with the dataset and are used to divide the dataset into context-sensitive groups. We also assume, in this work, that the variables are observed during test time and need not be inferred.

In order to learn models that can adapt to the subtle differences in the input-output mapping in different context-sensitive groups, we make use of a hierarchical Bayesian framework. Hierarchical models are suitable for problems where the data can be structured into groups, as such models allow the learning of parameters specific to each group while utilizing the entirety of the data [15]. Here, we adapt the hierarchical Bayesian neural network framework presented by Joshi et al. [17], who used it to model subject-specific gesture recognition, for our problem.

Instead of modeling individual subject-specific variances in gesture performance, we aim to capture the subtle context-sensitive group-specific variances in the input-expressivity mapping. We separate the training data into context-sensitive groups and train our hierarchical model using multimodal feature descriptors of each training video (Fig. 1). In order to predict the facial expressivity score from a test video, we use the parameters of the trained model associated with the context-sensitive group to which the test video belongs.

In summary, our contributions are: (1) we explore appropriate feature representations from multiple modalities (video and audio) to best predict facial expressivity in both multi-class classification and regression settings, (2) we compute the feature importance scores to investigate the relative importance of interpretable features in the task of expressivity prediction, and (3) we demonstrate the benefits of using a framework that adapts to contextual information.

II. RELATED WORK

Automatic analysis of facial expressions and affect has been an actively researched topic in the fields of computer vision and machine learning [12]. Many early works focused on the recognition of prototypic emotions from static images [22] or video [10]. A more detailed descriptor of the physical changes in the shape and texture of the face, named the Facial Action Coding System (FACS) was developed by Ekman and Friesen [13] to describe facial expressions in terms of anatomically defined Action Units (AUs). The problem of automatically identifying the presence [2] and intensity [21] of AUs from images [25] and video [7] has received a lot of attention in recent years. Progress in this field has led to development of several off-the-shelf applications [11], [3] capable of detecting AU presence and intensity values for several Action Units.

The dynamics of a person's face can provide information regarding the person's emotional state, intention and personality, as well as cognitive and biomedical status. The development of computational analyses techniques of facial expressions has opened avenues for researchers to view investigations of emotional and cognitive impairments using a computational lens. For example, Cohn et al. [9] conducted a feasibility study of detecting depression using facial actions and vocal prosody. Wang et al. [32] analysed video-based facial expressions to study neuropsychiatric disorders such as Asperger's Syndrome and Schizophrenia.

In the context of Parkinson's disease, Wu et al. [33] conducted a preliminary study to quantify facial expressivity of patients with PD by comparing AU activations between a group of 7 Parkinson's patients and 8 control patients. The authors quantify facial expressivity by manually defining a mathematical formula based on automatically detected AUs, and demonstrate a significant difference in facial expressivity between the control group and the patients. Bandini et al. [4] reported, from a pilot experiment involving 4 patients and 4 people in a control group, that control subjects exhibit higher distances from a neutral face when expressing emotions compared to PD patients. Almutiry et al. [1] found that certain expressions, such as happiness and disgust, are most discriminative when comparing the expressive behavior of PD patients with healthy controls.

In this work, our focus is not on quantifying differences in facial expressivity characteristics between PD patients and individuals in a control group. Instead, we use a larger dataset of approximately 800 data points of 117 patient interview audio-video inputs and their corresponding expertly annotated facial expressivity labels to automatically *learn* a function that maps the multimodal input feature representation to the facial expressivity score.

Most existing work on problems in affect and expression analysis focus on building generic and generalizable classifiers (e.g. [31], [18]). However, there have been some works focusing on personalization of classifiers, i.e. tailoring classifiers to adapt to individual variances, e.g. in modeling facial AU intensity [34], [7] and pain recognition [19].

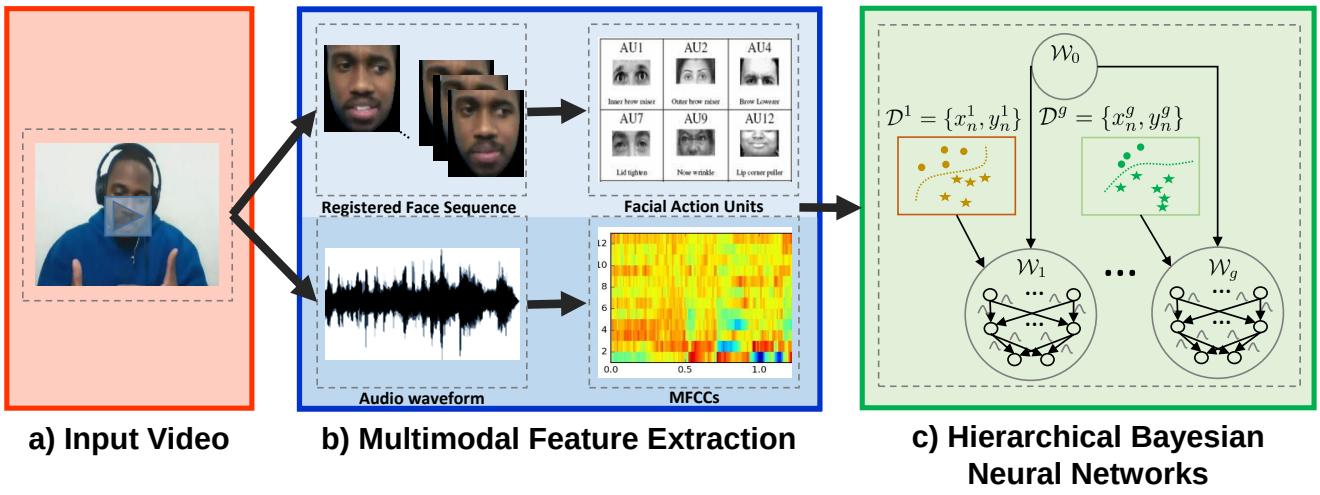

Fig. 2. Overview of our multimodal context-sensitive expressivity prediction model. From an input audio-video clip (a), we extract Facial Action Unit-based interpretable features as well as Mel-Frequency Cepstral Coefficient Features (b). We train a context-sensitive expressivity model by utilizing a hierarchical Bayesian neural network framework (c). Here $\mathscr{D}^1, ..., \mathscr{D}^g$ represents our dataset $\mathscr{D}$ divided into g context-sensitive groups, which we hypothesize to have subtly different input-label mappings. $\mathscr{W}_1, ..., \mathscr{W}_g$ represent the group-specific weights that parameterize the mapping between the input and the expressivity ratings.

Rudovic et al. [23] use a context-sensitive model to estimate AU intensity, where context is defined by *who*: the identity of the individual, *when*: the timing of the facial expressions and *how*: how the facial expressions change over time. Here, we utilize two definitions of context, *gender* and *sentiment*, to divide the training data into context-sensitive groups and train a context-sensitive hierarchical classification model.

III. SYSTEM OVERVIEW

A. Input

Fig. 2 summarizes our system. The input consists of short interview audio-video clips (approximately 20 seconds) of PD patients. The subjects are roughly front-facing with the face visible and well-illuminated and the audio is recorded clearly. Each video consists of the patient speaking about either some recent positive or negative experience. Associated with each of these videos is a facial expressivity rating, which we aim to predict.

B. Feature Extraction

For each frame of all the videos in the filtered dataset, 18 AU presence and 17 AU intensity values are extracted using OpenFace [3]. In order to explore whether facial expressivity can be predicted from the raw audio, we extract Mel-Frequence Cepstral Coefficient (MFFC) features using the Librosa library [20]. In order to compute an aggregate feature representation from the per-frame AU presence and intensity values, we use statistics (mean, standard deviation, min and max) for each feature to produce a 140-dimensional visual feature representation (AU-stats). We also compute the same statistics from the MFCC features to obtain a 160-dimensional audio feature representation (MFCC-stats). These feature descriptors serve as the input to both our hierarchical Bayesian neural network (HBNN-C) classification and hierarchical Bayesian neural network regression (HBNN-R) frameworks (Fig. 2b). The computed statistics enables the encoding of the range and amplitude of various audio-visual features along with the deviations from the mean.

C. Context Sensitivity

We experiment with two notions of context: gender (male and female) and sentiment (positive and negative) expressed in the interview. We wish to investigate whether dividing the dataset into context-sensitive groups and leveraging any variations inherent in the groups' input-label mapping can yield improvements in model performance. For example, previous research has indicated people display varying levels of expressive behavior when discussing positive experiences compared to when speaking about negative experiences [24]. Utilizing a framework that is capable of learning related but slightly different functions seems apt for such a scenario.

We assume we have access to context indicators, i.e. the subject's gender and the sentiment of the experience that the subject describes, for each video in both the training and test sets. This allows us to separate the dataset into context-specific groups (Fig. 2c).

D. Hierarchical Bayesian Neural Networks

Let $\mathscr{D} = \{x_n, y_n\}_{n=1}^N$ represent our dataset, containing N feature representations of N input audio-video clips $x_n \in \mathbb{R}^D$ of PD patients, and their corresponding facial expressivity ratings y_n. We wish to learn the functional mapping from input representations to expressivity ratings. We assume that $\mathscr{D}$ can be separated into G distinct context-sensitive groups.

We adopt, as the basis of our technique, the hierarchical framework introduced by Joshi et al. [17] and describe how it is applied for our problem. The conditional distributions of each context-sensitive group are parameterized via feed-forward neural networks, which can model subtle variations

in the input-expressivity mapping among groups (Fig. 3). Assuming the data instances are independent, we have,

$$p(\mathbf{y} \mid \mathscr{W}, \mathbf{c}, \mathbf{x}) = \prod_{n=1}^{N} \prod_{g=1}^{G} p(y_n \mid f(\mathscr{W}_g, x_n))^{\mathbf{1}[c_n=g]}. \quad (1)$$

Here, c_n is a categorical random variable representing which context-sensitive group x_n belongs to. In our experiments, we assume that the group indicators $\mathbf{c} = \{c_n\}_{n=1}^{N}$ are known during training and testing. We wish to learn $\mathscr{W} = \{\mathscr{W}_1, \ldots, \mathscr{W}_G\}$, where $\mathscr{W}_g$ is the set of group-specific weights parameterizing either a neural network classifier or a neural network regressor f.

For a neural architecture with 1 hidden layer, we have,

$$h = ReLU(w^g_{l=0} x), \quad (2)$$

$$f = \mathscr{S}(w^g_{l=1} h), \quad (3)$$

where, $\mathscr{S}(a) = \exp\{a\}/\sum_k \exp\{a_k\}$ is the softmax function, ReLU (Rectified Linear Unit) represents a non-linear activation function, $w^g_{l=0}, w^g_{l=1}$ represents the weights of layer 0 and 1 respectively, x represents the input and h represents an intermediate hidden representation. Note that the softmax function is not required when training a regression network.

As in [17], factorized Gaussian priors are placed on $\mathscr{W}_g$ with independent group-specific variances. This models our prior assumption that each context-sensitive group's functional input-output mapping is an independently corrupted version of some common latent mapping, $\mathscr{W}_0$,

$$p(\mathscr{W}_g \mid \mathscr{W}_0, \tau_g) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathscr{N}(w^g_{ij,l} \mid w^0_{ij,l}, \tau_g^{-1}). \quad (4)$$

Uninformative priors, zero mean Gaussians with a large fixed variance τ_0^{-1}, are placed on the weight means $\mathscr{W}_0$,

$$p(\mathscr{W}_0 \mid \tau_0) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathscr{N}(w^0_{ij,l} \mid 0, \tau_0^{-1}). \quad (5)$$

Here, V_l represents the number of units in layer l, with $l = 0$ corresponding to the input layer of the neural network.

In order to infer the group-specific standard deviations from the data, we use the half-normal distribution with a large fixed variance v as hyper-priors over the group-specific standard deviations $\tau_g^{-1/2}$,

$$p(\gamma_g \mid v) = \mathscr{N}(\gamma_g \mid 0, v); \quad \tau_g^{-1/2} = |\gamma_g|, \quad (6)$$

where an auxiliary variable γ_g has been used. If $a \sim \mathscr{N}(0, \sigma^2)$, then $|a| \sim \text{Half-Normal}(0, \sigma^2)$. For the classification network, the observed class labels are modeled as categorically distributed random variables,

$$y_n \mid \mathscr{W}, x_n, c_n \sim \text{Cat}(y_n \mid f(\mathscr{W}_{c_n}, x_n)). \quad (7)$$

That is, y_n represents the probability distribution over the facial expressivity classes. For the regression network, the observed labels are modeled as normally distributed random variables,

$$y_n \mid \mathscr{W}, x_n, c_n \sim \mathscr{N}(y_n \mid f(\mathscr{W}_{c_n}, x_n)). \quad (8)$$

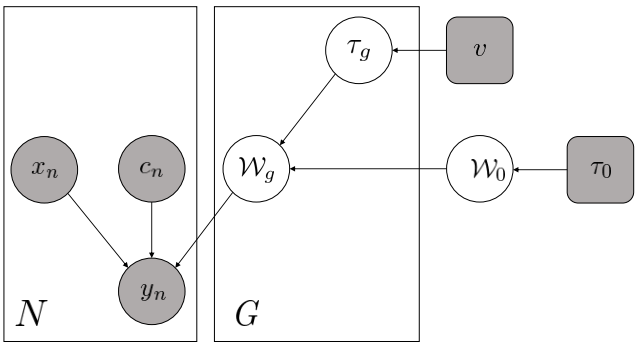

Fig. 3. Graphical representation of our hierarchical Bayesian framework. x_n, y_n represent the input-output pair, while c_n indicates the context-sensitive group-membership of data sample n. $\mathscr{W}_g$ represents the set of group-specific weights parameterizing a Bayesian neural network f. Shaded nodes indicate that the random variables are observed.

We can summarize the joint distribution specified by the model as,

$$\begin{aligned} p(\mathscr{W}_0, \mathscr{W}, \mathscr{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{c}, \tau_0, v) = p(\mathscr{W}_0 \mid \tau_0^{-1}) \\ \prod_{g=1}^{G} p(\gamma_g \mid v) p(\mathscr{W}_g \mid \mathscr{W}_0, \tau_g^{-1}) \\ \prod_{n=1}^{N} \prod_{g=1}^{G} p(y_n \mid f(\mathscr{W}_g, x_n))^{\mathbf{1}[c_n=g]}, \end{aligned} \quad (9)$$

where $\mathscr{T} = \{\gamma_1, \ldots, \gamma_G\}$. The hierarchical Bayesian neural network learns the context-sensitive group-specific variances by allowing the group-specific conditional distribution of data from different groups to vary from each other, while allowing the sharing of statistical strength across groups.

E. Learning

Because the posterior distribution cannot be learned analytically, we use variational inference to learn a tractable approximation,

$$q(\mathscr{W}_0, \mathscr{W}, \mathscr{T} \mid \phi) = q(\mathscr{W}_0 \mid \phi_0) \prod_{g=1}^{G} q(\mathscr{W}_g \mid \phi_g) q(\gamma_g \mid \phi_{\gamma_g}), \quad (10)$$

where $\phi = \{\phi_0, \phi_1, \ldots, \phi_G, \phi_{\gamma_1}, \ldots, \phi_{\gamma_G}\}$ represents the variational free parameters. The weight posteriors are approximated with fully factorized Gaussian distributions. The variational parameters are optimized by minimizing the Kullback-Liebler divergence $\text{KL}(q||p)$ between the true posterior and the variational approximation by maximizing the expected lower bound (ELBO),

$$\begin{aligned} \mathscr{L}(\phi) = \mathbb{E}_{q_\phi}[\ln\ p(\mathscr{W}_0, \mathscr{W}, \mathscr{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \gamma_0, v)] \\ - \mathbb{E}_{q_\phi}[\ln\ q(\mathscr{W}_0, \mathscr{W}, \mathscr{T} \mid \phi)], \end{aligned} \quad (11)$$

with respect to the variational free parameters ϕ.

Given a test video, the posterior predictive distribution is evaluated by a Monte-Carlo estimate using the optimal variational parameters corresponding to the group to which the test video belongs.

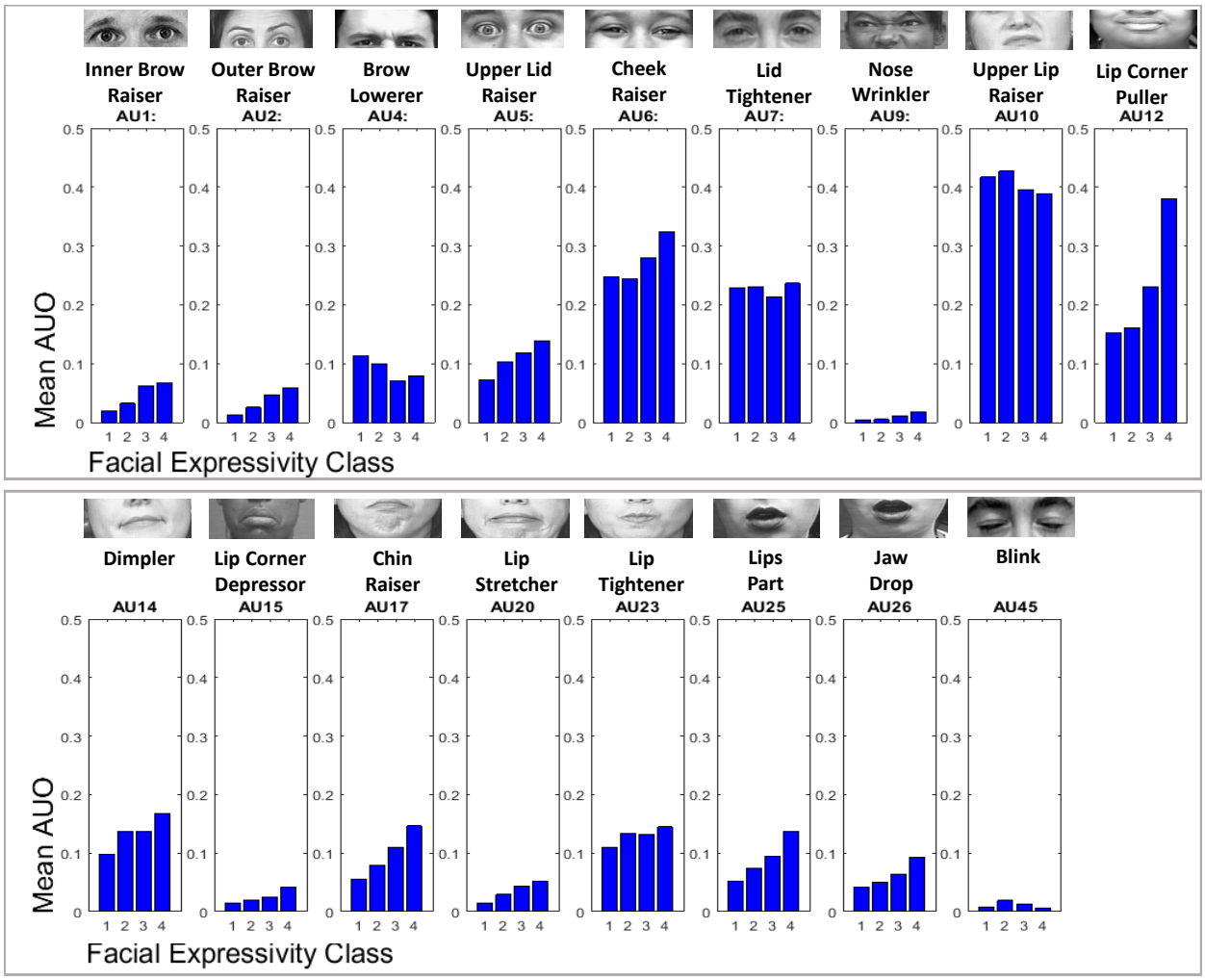

Fig. 4. For each class in the dataset, the average Action Unit Occurence (AUO) is plotted for 17 different AUs. For each subplot, the x-axis represents the 4 facial expressivity classes whereas the y-axis represents the mean AUO score as defined in Eq. 11. Grayscale images depicting the AUs were obtained from https://www.cs.cmu.edu/∼face/facs.htm

IV. EXPERIMENTS

A. DATASET

We use the dataset that was originally collected in a previous study to investigate the effects of self-management rehabilitation on health-related quality-of-living in Parkinson's disease [27]. All 117 participants in this study had been diagnosed with PD and had the ability to understand and to communicate with personnel [27]. Patients were videotaped during social interactions with occupational therapy practitioners with the camera recording a frontal face view.

From the videotapes of the interactions, 20-second representative segments of patients speaking about a positive and negative experience were extracted for analysis. Each of the extracted videoclips was given 5-point Likert scale ratings for the different variables of the ICRP, one of which corresponds to active expressivity of the face, by at least four trained researchers. Because multiple annotators labeled each video, a composite score was computed by taking the mean of the scores provided by each rater. Using the intra-class correlation coefficient (ICC), the inter-rater reliability for the variable representing the active expressivity in the face was reported to be .89 (for n = 4 raters) and .67 (for n = 1 rater) [26], suggesting a reasonable level of agreement when trained raters code this attribute.

Data Analysis: For our experiments, video samples where the subject's face could not be detected in a sufficient number (30) of frames due to occlusion or bad illumination were discarded while building our expressivity prediction model, reducing the size of the dataset from 805 to 772 video samples. This threshold was chosen to maintain a sufficient number of video-label pairs for training while ensuring that features could be extracted from each sample in order to contribute to building an accurate model. The ground truth expressivity labels $y_i \in \mathbb{R}$ for each video is taken as the average of 4 expert ratings. In our experiments, we evaluate both regression and classification formulations in predicting the expert ratings. For classification experiments, we discretize the labels of the entire dataset into 4 classes. Classes 1, 2, 3 and 4 contain samples with facial expressivity

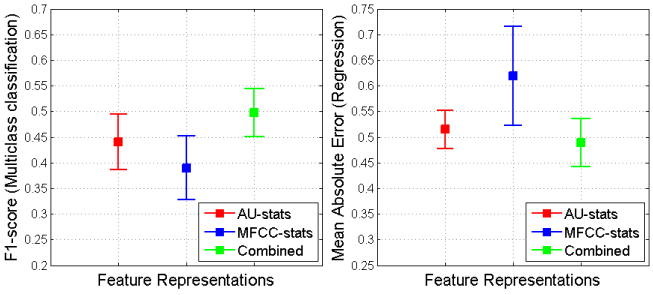

Fig. 5. (Left) The mean F1-scores along with their standard deviations for HBNN-C models trained with different feature representations. (Right) The mean MAE scores along with their standard deviations for HBNN-R models trained with different feature representations. Note that MAE is a measure of the absolute difference between the model predictions and the ground truth, which ranged from 1 to 5. Using a combined multimodal feature representation of both video and audio features yielded the best performance in both the classification (mean F1-score of 0.50) and regression (mean MAE-score of 0.49) settings.

ratings in the range [1, 2), [2, 3), [3, 4) and [4, 5] respectively.

B. FACIAL ACTION UNIT ANALYSIS

Because AUs are precisely defined, they serve as good candidates to use as features in applications requiring interpretability. We visualized how often and with what intensity various action units occur on average for the different expressivity classes of the entire dataset. For each video, we aggregated AU presence values weighted by their respective intensity values and normalized them by the total number of frames in which the face was detected:

$$AUO_a = \frac{1}{N}\sum_j AUI_a^j \times AUP_a^j, \tag{12}$$

where, AUO_a represents the mean Action Unit Occurence for AU a of the video, N represents the number of frames in the video in which the face was detected and AUI_a^j and AUP_a^j represent the presence and intensity values of AU_a for frame j respectively.

For all videos belonging to a specific facial expressivity class, we computed the mean AUO for 17 AUs whose presence and intensities were detected by OpenFace and plot them (Fig. 4). Although it is challenging for automatic AU recognition methods to generalize well to datasets beyond the ones on which the models have been trained, we observed that they capture enough signal allowing the analysis of some interesting trends.

On average, we found that the $AUOs$ for several AUs increased when facial expressivity increases. For example, AUs corresponding to brow raising (AU1, AU2), lip corner pulling (AU12), chin raising (AU17), lip stretching (AU20) and jaw dropping (AU26) occured more frequently in videos with higher expressivity values. This indicates that in people deemed to have higher values of facial expressivity, certain Action Units are more frequently activated with higher intensities. For other action units, such as AUs corresponding to brow lowering (AU4), lip tightening (AU23) and blinking (AU45), a clear linear trend was absent. The AU representing upper lip raising (AU10) has the highest AUO values across all classes on average due to the fact that patients are speaking for the entire duration of the video clip.

TABLE I
ACTION UNIT FEATURE IMPORTANCE SCORES

Action Unit	Feature Importance
AU5 (Upper Lid Raiser)	0.088
AU12 (Lip Corner Puller)	0.081
AU25 (Lips Part)	0.075
AU26 (Jaw Drop)	0.065
AU4 (Brow Lowerer)	0.059
AU14 (Dimpler)	0.052
AU10 (Upper Lip Raiser)	0.048
AU23 (Lip Tightener)	0.044
AU2 (Outer Brow Raiser)	0.039
AU1 (Inner Brow Raiser)	0.037
AU7 (Lid Tightener)	0.027
AU9 (Nose Wrinkler)	0.026
AU15 (Lip Corner Depressor)	0.024
AU6 (Cheek Raiser)	0.004
AU20 (Lip Stretcher)	0.001
AU17 (Chin Raiser)	-0.012
AU45 (Blink)	-0.028

C. FACIAL EXPRESSIVITY PREDICTION

We trained classification models (HBNN-C) for the multiclass setting, where classes were computed from labels as described in Section IV, and regression models (HBNN-R) using the original expressivity ratings. We performed subject-independent 9-fold cross-validation for all our experiments. For both regression and classification experiments, we trained a model with 1 hidden layer containing 50 hidden nodes. The hyper-parameters ν and τ_0 of the model were set to 100 and 1000 respectively and RMSprop [29] was used for optimization.

Multimodality: We trained HBNN-C and HBNN-R models with all data pooled into one group and experimented with different feature representations as input to our classifier (Fig. 5). We trained our model with visual features, consisting of Action Unit statistics (AU-stats), audio features, consisting of MFCC statistics (MFCC-stats), as well as a combined audio-video feature representation (Combined).

We found that the model trained solely on AU-stats obtained a mean F1-score of 0.44 in the multiclass classification setting and a mean absolute error (MAE) of 0.51 in the regression setting. Multiway classification is challenging in this scenario with neighboring classes often getting confused with each other. However, this is to be expected because even expert human raters only agree unanimously in their Likert-scale labels 25.6% percent of the time for this dataset.

We found that using featues computed from the raw audio also led to reasonable model performance (0.39 mean F1-score and 0.62 mean MAE in the classification and regression settings respectively). In instances where the video is missing, corrupted, or of low quality for automated facial analysis, expressivity can be estimated solely from audio. Using a combined multimodal feature representation of both

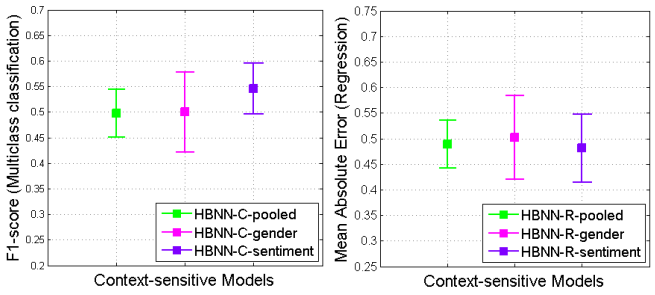

Fig. 6. (Left) The mean F1-scores and their standard deviations for HBNN-C models trained using context (gender, sentiment) or no context (pooled). (Right) The mean MAE scores and their standard deviations for HBNN-R models trained using context (gender, sentiment) or no context (pooled).

video and audio features yielded the best performance in both the classification (mean F1-score of 0.50) and regression (mean MAE-score of 0.49) settings.

Feature Importance: Feature importance scores indicate how useful a given feature or attribute is in the classification task. One simple and interpretable method of estimating the importance score for a feature is to measure the difference in F1-scores before and after randomly permuting the values of the feature during training [6]. If the difference is large, one can assume that it plays an important role in classification whereas if the difference is negligible, one can assume that the feature has little importance.

Using the classification model trained on visual features (AU-stats), we computed an estimate of feature importance for all visual features (AU-stats) over all folds. In order to obtain a heuristic of importance associated with each individual AU for the task of facial expressivity prediction, we averaged the scores for all features associated with any given action unit. For example, the importance score for AU5 (Upper Lid Raiser) is computed by taking the mean of all feature importance scores corresponding to the 8 features associated with AU5 (means, standard deviations, maxes and mins of AU5 presence and intensity values). The action units sorted in order of their estimated importance scores are presented in Table 1.

AU5 (Upper Lid Raiser), often associated with expressions displaying shock and anger, and AU12 (Lip Corner Puller), associated with expressions containing smiles, scored the highest in the AU importance heuristic, whereas AU45 (Blink) scored the lowest. It is interesting to note that *AUO*s computed for AU5 and AU12 (Fig. 4) exhibited an increasing trend with higher expressivity, which is absent for AU45.

Context-sensitive Classification: We experimented with two different notions of context: gender and the sentiment expressed in the patient interviews. For each context indicator, we first divided the training data into two groups (male and female for gender, positive and negative for sentiment). We trained our framework using this multi-group paradigm with the combined audio-video feature representation. During testing, we obtained the estimate of the expressivity rating of the test sample using the classification or regression parameters associated with its corresponding context indicator.

Compared to a baseline model (HBNN-C-pooled), which ignored context and was trained with the data from all groups pooled together (obtaining a mean F1-score of 0.50), we found that retaining contextual information provided by gender (HBNN-C-gender) yielded no empirical benefit in classifier performance (mean F1-score of 0.50). However, utilizing the context provided by sentiment (HBNN-C-sentiment) improved the performance of the model in the multiclass classification settings (mean F1-score of 0.55) (Fig. 6). Similarly, the regression model that utilized context provided by sentiment (HBNN-R-sentiment) yielded a slightly improved MAE score of 0.48, outperforming the baseline model which obtained a mean MAE score of 0.49 (Fig. 6). This suggests that the input-label mappings in the sentiment-driven context-sensitive groups may contain sufficient group-specific variance in order for the hierarchical framework to leverage it into improved model performance.

V. CONCLUSIONS

Automated assessment of facial expressivity in Parkinson's Disease patients has the potential to be a useful tool for clinicians in this field. However, most existing works in the domain are limited to small-scale pilot studies comparing the characteristics and dynamics of facial expressions exhibited by a small group of PD patients and a separate control group. In this work, we utilized a dataset of 772 short audio-video clips of 117 PD patients along with their facial expressivity labels to train a machine learning model capable of predicting the facial expressivity ratings of new audio-video clips.

We provided an analysis of facial expressivity in terms of how often various facial Action Units are activated in the videos in the dataset, weighing the activations by their intensities. We observed an increasing trend of AU occurences for several action units, such as AUs 1, 2, 5 and 12 with increasing facial expressivity. We also computed a heuristic importance score for each AU and found that AUs 5 and 12 were deemed most important in the expressivity prediction task, while AUs 17 and 45 were deemed the least important.

We demonstrated the utility of extracting features from not only the visual domain but also the audio in order to accurately predict facial expressivity, finding that a model trained on a combined audio-visual feature representation outperformed models trained on features extracted from a single modality for both classification and regression tasks.

Finally, we illustrated the benefits of using a framework that adapts to contextual information. Our hierarchical Bayesian model trained on a dataset divided according to the sentiment expressed in the interviews outperformed a baseline model that ignored this contextual information in both classification and regression scenarios.

Future extensions of this work include personalizing the contextual model further to individuals and developing a mechanism for the model to automatically determine contextual clusters in the training data in the absence of pre-defined context labels.

REFERENCES

[1] R. Almutiry, S. Couth, E. Poliakoff, S. Kotz, M. Silverdale, and T. Cootes. Facial behaviour analysis in parkinsons disease. In *International Conference on Medical Imaging and Virtual Reality*, pages 329–339. Springer, 2016.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[4] A. Bandini, S. Orlandi, H. J. Escalante, F. Giovannelli, M. Cincotta, C. A. Reyes-Garcia, P. Vanni, G. Zaccara, and C. Manfredi. Analysis of facial expressions in parkinson's disease through video-based automatic methods. *Journal of Neuroscience Methods*, 281:7–20, 2017.

[5] M. Bologna, G. Fabbrini, L. Marsili, G. Defazio, P. D. Thompson, and A. Berardelli. Facial bradykinesia. *J Neurol Neurosurg Psychiatry*, 84(6):681–685, 2013.

[6] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[8] J. F. Cohn and F. De la Torre. Automated face analysis for affective. *The Oxford handbook of affective computing*, page 131, 2014.

[9] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009.

[10] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference On*, pages 396–401. IEEE, 1998.

[11] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. Intraface. In *Automatic Face and Gesture Recognition*, 2015.

[12] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual analysis of humans*, pages 377–409. Springer, 2011.

[13] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[14] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[15] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[16] S. D. Gunnery, E. N. Naumova, M. Saint-Hilaire, and L. Tickle-Degnen. Mapping spontaneous facial expression in people with parkinsons disease: a multiple case study design. *Cogent Psychology*, page 1376425, 2017.

[17] A. Joshi, S. Ghosh, M. Betke, S. Sclaroff, and H. Pfister. Personalizing gesture recognition using hierarchical bayesian neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. In *European conference on computer vision*, pages 649–662. Springer, 2010.

[19] D. Lopez-Martinez and R. Picard. Multi-task neural networks for personalized pain recognition from physiological signals. *arXiv preprint arXiv:1708.08755*, 2017.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[21] J. Nicolle, K. Bailly, and M. Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

[22] M. Pantic and L. J. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[23] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):944–958, 2015.

[24] K. Takahashi, L. Tickle-Degnen, W. J. Coster, and N. K. Latham. Expressive behavior in parkinsons disease as a function of interview context. *American Journal of Occupational Therapy*, 64(3):484–495, 2010.

[25] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

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

[27] L. Tickle-Degnen, T. Ellis, M. H. Saint-Hilaire, C. A. Thomas, and R. C. Wagenaar. Self-management rehabilitation and health-related quality of life in parkinson's disease: A randomized controlled trial. *Movement Disorders*, 25(2):194–204, 2010.

[28] L. Tickle-Degnen and K. D. Lyons. Practitioners impressions of patients with parkinson's disease: the social ecology of the expressive mask. *Social Science & Medicine*, 58(3):603–614, 2004.

[29] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[30] M. Valstar. Automatic facial expression analysis. In *Understanding Facial Expressions in Communication*, pages 143–172. Springer, 2015.

[31] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43, 2012.

[32] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of neuroscience methods*, 168(1):224–238, 2008.

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

[34] S. Yang, O. Rudovic, V. Pavlovic, and M. Pantic. Personalized modeling of facial action unit intensity. In *International Symposium on Visual Computing*, pages 269–281. Springer, 2014.