2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

Beyond FACS: Data-driven Facial Expression Dictionaries, with Application to Predicting Autism

Evangelos Sariyanidi,[1] Lisa Yankowitz,[1] Robert T. Schultz,[1,2] John D. Herrington,[1,2]
Birkan Tunc[†1,2] and Jeffrey Cohn[†3]
[1] Center for Autism Research, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
[2] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA
[3] Deliberate AI, New York, NY, USA.
[†] Equal contribution.

***Abstract*— The Facial Action Coding System (FACS) has been used by numerous studies to investigate the links between facial behavior and mental health. The laborious and costly process of FACS coding has motivated the development of machine learning frameworks for Action Unit (AU) detection. Despite intense efforts spanning three decades, the detection accuracy for many AUs is considered to be below the threshold needed for behavioral research. Also, many AUs are excluded altogether, making it impossible to fulfill the ultimate goal of FACS—the representation of *any* facial expression in its entirety. This paper considers an alternative approach. Instead of creating automated tools that mimic FACS experts, we propose to use a new coding system that mimics the key properties of FACS. Specifically, we construct a data-driven coding system called the Facial Basis, which contains units that correspond to localized and interpretable 3D facial movements, and overcomes three structural limitations of automated FACS coding. First, the proposed method is completely unsupervised, bypassing costly, laborious and variable manual annotation. Second, Facial Basis reconstructs all observable movement, rather than relying on a limited repertoire of recognizable movements (as in automated FACS). Finally, the Facial Basis units are additive, whereas AUs may fail detection when they appear in a non-additive combination. The proposed method outperforms the most frequently used AU detector in predicting autism diagnosis from in-person and remote conversations, highlighting the importance of encoding facial behavior comprehensively. To our knowledge, Facial Basis is the first alternative to FACS for deconstructing facial expressions in videos into localized movements. We provide an open source implementation of the method at `github.com/sariyanidi/FacialBasis`.**

I. INTRODUCTION

Since its initial development more than four decades ago [1], the Facial Action Coding System (FACS) has been widely used to study how facial expressions relate to emotion, personality, mood, deception, and mental health [2]. With the advent of computer vision, researchers have aimed to automate the detection of FACS Action Units (AUs), as manual AU coding is laborious, costly, and requires extensive training [3]. However, despite nearly three decades of research, the accuracy of automated systems is often below the threshold needed for behavioral research [4]. Moreover, available AU detection software excludes a large number of AUs—and even the AUs that are included may fail detection if they are annotated with low reliability or appear in a non-additive AU combination (Section II-A). As such, automated FACS coders behave more akin to a retrieval system that checks for the presence of certain AUs or AU combinations, rather than fulfilling the original purpose of FACS, which is encoding any facial expression in its entirety. The consequences of this major goalpost shift remain unknown and are potentially severe for progress in behavioral and medical sciences. The inability of automated FACS coders to fully encode expressions can prevent the discovery of behavioral patterns that characterize the full range of emotions, personality traits or mental health conditions.

Instead of developing a new automated FACS coder, we propose a new, alternative coding system that is more amenable to automated expression measurement. Specifically, we propose to construct a data-driven coding system by learning a dictionary [5], which can overcome three significant inherent limitations of automated FACS coding. First, dictionaries reconstruct all observable facial movements, therefore can encode expressions comprehensively, in contrast with FACS software that provide results only for the AUs in their repertoire. Second, dictionaries are learned in an unsupervised manner. This is a significant advantage over FACS-based approaches, as the costly and laborious manual annotation needed by the latter may render large parts of available videos unusable for supervised training (Section II-A). Third, all basic expression units in a dictionary are additive—they can be detected successfully in isolation or in combination with other expression units (Section II-A).

To show the utility of data-driven dictionaries in clinical applications, we compare multiple facial coding systems, including (automated) FACS coding, in classifying adolescents with autism (AUT) vs. those who are neurotypical (NT). We experiment on two datasets of naturalistic conversational tasks: one with in-person conversations, and one with remote conversations. Studying both contexts allows us to assess whether behavioral symptoms that characterize autism can be effectively measured from remote conversations, which is needed given the post-pandemic increase in telehealth assessment of autism [6]. The present results suggest that the proposed system outperforms the most widely used

This work is partially supported by the Office of the Director (OD), National Institute of Child Health and Human Development (NICHD), and National Institute of Mental Health (NIMH) of US, under grants R01MH118327, R01MH122599, 5P50HD105354-02 and R21HD102078; and the IDDRC at CHOP/Penn.

979-8-3315-5341-8/25/$31.00

automated FACS coder, namely OpenFace [7], in AUT vs. NT classification both on the in-person and the remote sample. Furthermore, results from remote conversations underpin the importance of comprehensively encoding facial expressions (Section IV-C.2), suggesting that the exclusion of many AUs from automated pipelines (Section II-A) may be consequential.

Our findings indicate that developing data-driven coding systems that retain the advantages of FACS, rather than automating FACS, is a potent new paradigm for empowering mental health research (Section VI), since the structural limitations of automated FACS coding (Section II-A) may not be overcome by developing more sophisticated AU detectors. We provide an end-to-end open-source toolkit[1] that can be used for conducting behavioral research with the Facial Basis. To our knowledge, this is the first open-source software that provides an alternative to automated FACS for quantifying facial expressions in 2D videos by breaking them down into localized expression units.

In sum, the contributions of this paper are as follows. We:

- Show that data-driven dictionaries are a viable alternative to FACS for supporting mental health research with an interpretable expression coding system.
- Experiment using in-person and remote conversations and show that autism can be predicted in both contexts.
- Show that the behavioral symptoms that are most predictive of AUT vary between contexts.
- Provide an open-source toolkit that to our knowledge contains the first end-to-end software pipeline for producing localized expression coefficients from a data-driven coding system.

II. Related Work

A. Automated Action Unit Detection

Automated FACS coding has long been a subject of intense research in computer vision [8]–[10], as it can support a variety of industrial and research applications [11]. However, there is still a need for improving accuracy, as the average F1 score of even state-of-the-art AU detectors is in the range of 0.60-0.67 [7], [12]–[16]. The accuracy on cross-database experiments, which are indicative of real-world performance, tends to be even lower and below the threshold needed for reliable behavioral research [4]. In an era where AI algorithms deliver impressive results across domains, the limited progress in automating FACS can be attributed to four structural barriers.

First, improving accuracy by increasing the training data is difficult. Training supervised models necessitates manual AU annotation, which is laborious and requires multiple trained experts. Even when longer video recordings of naturalistic social interactions are available, researchers are usually restricted to use only a few minutes or just seconds long segments (*e.g.*, most facially-expressive 20 seconds [17]) due to the infeasibility of longer annotations. Second, AU labels are usable only if they pass a certain level of inter-rater reliability [1] which can be low for certain AUs [17], [18], reducing the usable video data further. Third, reliability tends to drift over time and between independent coders. Fourth, AUs that have a low base rate are difficult to detect with a supervised classifier may be excluded from automated pipelines altogether [16]. For example, existing toolkits provide outputs for only 17-19 AUs [7], [16], [19], whereas the original FACS contains 45 AUs (30 for the revised version) [20]. Finally, the AUs are generally not additive, as they can modify each other's appearance. A typical example is AU 1+4 [21]. When AU 1 occurs alone, the inner eyebrows are pulled upward. When AU 4 occurs alone, they are pulled together and downward. When AU 1 and AU 4 occur together, they result in appearance changes that do not occur in either AU 1 or AU 4 in isolation: The inner eyebrows are raised and pulled together, giving the brows an oblique shape, and causing wrinkles to appear in the center of the forehead. The existence of non-additive AUs suggest that automated FACS coders can comprehensively encode all expressions only if they are trained with datasets that contain all isolated AUs as well as non-additive AU combinations, which practically is not possible.

While FACS is a powerful coding system, the above-listed barriers significantly restrict the upper limit of accuracy achievable by automated FACS coders. Our study focuses on developing a new coding system that retains a key property of FACS – breaking down expressions into localized units – while being more amenable to automation.

B. Coding Systems Based on Linear Models

Linear decomposition based on sparse dictionary learning [5] provides an alternative coding system that can readily overcome a fundamental limitation of FACS, namely, the lack of additivity (Section III-B). In particular, sparse dictionaries can be trained to contain elements that correspond to localized expression units similar to AUs. An implementation of this idea for facial expression analysis is the Facial Bases method [22]. However, this method is based on a 2D pixel representation where separating out-of-plane head motion from facial expressions is not feasible. For this reason, producing meaningful expression coefficients from interactions that involve head movements (*e.g.*, conversations) is very challenging. Moreover, the Facial Bases method has not been compared to automated FACS in a clinical context, thus its ability to serve as an alternative coding system for mental health research is yet to be shown.

A potential alternative is to use 3D morphable model (3DMM) fitting [23] for reconstructing the 3D face shape from 2D data, as expressions can be separated from head pose and identity in the 3D coordinate space [24]. A 3DMM typically contains an expression model (*e.g.*, FaceWarehouse [25]) that, in principle, may serve as a coding system. The expression models of 3DMMs, however, typically contain deformations that can be physically implausible and impossible to interpret, as they are often generated using global models such as PCA (Fig. 1; see Section III-B). In

[1] https://github.com/sariyanidi/FacialBasis

this paper, we define a sparse dictionary learning procedure that learns localized 3D facial expression units to generate physically plausible and interpretable movements akin to AUs (Section IV-C.1).

Similar interpretable and local linear models which operate in the 3D coordinate space have previously been used in the animation industry, where they are usually referred to as blendshapes [26]. The focus of blendshapes has been on video synthesis (3D to 2D) rather than analysis via reconstruction from videos. The number and content of the expression units in the blendshapes is determined according to this priority, which is not necessarily in line with the priorities of expression analysis. For example, the inclusion of as many as 946 expression units [26] may be warranted to generate person-specific differences in the appearance of expressions. Behavioral analysis, on the other hand, often demands a level of abstraction that ignores identity- or age-related differences in the generation of expressions. Representing facial behavior with hundreds of expression units can create multiple comparisons problems [27] and lead to multi-determined expression quantifications (same expression being represented by different components). As the distinctions between different expression units become too nuanced and difficult to semantically describe, reproducibility of quantitative findings becomes less attainable.

C. Predicting Autism from Conversations

Social communication is a core domain of impairment in autism [28]. A number of studies aimed to delineate communication differences between autistic and neurotypical participants using video recordings of conversations and computer vision tools, showing that the two groups can be successfully classified [29]–[31].

The most common approach adopted by autism researchers to quantify facial expressions has been automated AU detection, as FACS is an established coding system used for nearly five decades in behavioral sciences. In particular, OpenFace has been the most widely used software, used in a large number of autism studies [29], [32]–[39]. Although FACS is a reasonable choice in behavioral and medical sciences due to its interpretability and its ability to represent any possible facial expression, it is important to note that automated FACS coding is not equivalent to manual FACS coding. Detection accuracy for certain AUs can be low in automated software, and many AUs are entirely excluded, making it impossible to encode expressions comprehensively (Section II-A). To our knowledge, we apply the first alternative coding system to predict autism from conversational videos. Experiments show that the proposed system outperforms OpenFace in predicting autism, providing a promising alternative for studying mental health conditions.

A common paradigm in the literature involves unstructured or semi-structured conversations, where participants engage in face-to-face, in-person interactions with study staff [31]. The COVID-19 pandemic accelerated the use of remote videoconferencing for telehealth applications, including autism assessment, due to their scalability, ecological validity, and convenience [6]. It is therefore increasingly important to determine whether remote data collection paradigms can be effective in studying autism-related differences in social communication. To our knowledge, this is the first study that simultaneously conducts controlled AUT vs. NT classification using automated expression detection on data collected from the same paradigm from both in-person and remote conversations (Section IV-A). Results lead to two novel findings. First, in-person and remote conversations contain measurable differences in facial behavior between AUT and NT individuals, evidenced by a high and comparable classification accuracy (Section IV-C.2). Second, the behaviors that lead to highest classification accuracy are not necessarily identical between the two contexts.

III. Proposed Expression Coding System

The proposed coding system, Facial Basis, is a linear model that is comprised of localized expression components called the *Basis Units* (BUs). Each BU represents a localized movement in the 3D coordinate space. This coding system provides two advantages. First, a facial expression is represented as the linear sum of BUs; therefore, the coding system does not suffer from the existence of non-additive units. Second, operating in the 3D rather than the 2D space allows facial expressions to be separated from head pose- or identity-related variations in an image.

Below we describe how we reconstruct the 3D face shape and expression variation in a given 2D image via 3DMMs (Section III-A), and how we construct a coding system that represents a 3D expression variation as a linear sum of localized and semantically interpretable expression components (Section III-B).

A. Recovering Expression via 3DMM Fitting

Let $\mathbf{X} \in \mathbb{R}^{3P}$ be a vector that represents the 3D shape of a face that appears in a given image—a vector containing the 3D coordinates of P points *w.r.t.* the camera. Then, 3DMM fitting aims to reconstruct $\mathbf{X}$ as [24]

$$\mathbf{X} = \mathbf{R}(\bar{\mathbf{X}} + \mathbf{A}\boldsymbol{\alpha} + \mathbf{E}\boldsymbol{\varepsilon}) + \boldsymbol{\tau}, \tag{1}$$

where $\bar{\mathbf{X}}$ is the mean face of the 3DMM, and $\mathbf{A}$ and $\mathbf{E}$ are matrices that respectively represent the identity and expression models of the 3DMM. $\mathbf{R}$ and $\boldsymbol{\tau}$ account for the pose (rotation and translation) of the 3D face *w.r.t.* the camera, whereas the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ respectively represent the identity- and expression-related variation in the 3D face shape. The decoupling of facial expression from pose and identity as in the right-hand-side of (1) allows one to overcome two major challenges in expression analysis, namely, sensitivity to person-specific facial morphology (*i.e.*, identity bias) and to head pose [11]. It must be noted, however, that the decoupling of expression from pose or identity is an active research problem, and the degree to which it can be accurately accomplished depends on a number of factors, including the availability of the camera parameters [40] or to the usage of single or multiple frames while reconstructing identity [24], as well as the 3DMM fitting technique that is

Fig. 1. The first five components of the expression model used for 3DMM fitting [25]. Because this model is based on PCA, the components apply to the entire face and may be physically implausible.

used. Nevertheless, the coding system that we propose is not tied to a specific 3DMM fitting procedure; as long as the mesh topology that underlies the 3DMM is compatible, the same coding system can be used with any fitting procedure. This modular design ensures the relevance of the solution to future advances in methods for characterizing expression, pose and identity.

B. Localized Facial Basis

Once a 3DMM is fit to an image, the facial expression in the image can be represented as $\mathbf{E}\boldsymbol{\varepsilon}$ in (1), which is a $3P$-dimensional vector that describes how the 3D face shape deviates from the neutral face of the person. The matrix $\mathbf{E}$ can be considered as an expression coding system, since it explains the expression variation as a linear sum of its components (*i.e.*, columns) $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M$ as,

$$\mathbf{E}\boldsymbol{\varepsilon} = \varepsilon_1\mathbf{e}_1 + \varepsilon_2\mathbf{e}_2 + \cdots + \varepsilon_M\mathbf{e}_M. \tag{2}$$

That is, each of the components $\mathbf{e}_i$ can be considered as an expression code (*i.e.*, unit), and the coefficient ε_i indicates whether the expression variation encoded by $\mathbf{e}_i$ is present in the input image, while the magnitude of ε_i quantifies its intensity. However, the components $\{\mathbf{e}_i\}_i$ typically correspond to deformations that govern the entire face and can be physically implausible (Fig. 1), as they are learned with PCA or a similar global transformation. Moreover, usually all the expression coefficients $\{\varepsilon_i\}_i$ are activated for any image that contains an expression variation. These characteristics are in contrast with FACS, which can represent any expression with a small subset of localized and specialized AUs. For example, the prototypical expressions of the six basic emotions involve 2-7 AUs [41]. Indeed, the sparse activation pattern of FACS is a critical property for behavioral research, as it allows scientists to investigate which facial movements are related with an emotion, personality trait, or a mental condition.

Our coding system aims to reconstruct the expression $\mathbf{E}\boldsymbol{\varepsilon}$ with another linear model $\mathbf{W}$ that emulates two properties of FACS: Containing localized expression components (*i.e.*, columns), and having a sparse activation pattern. Thus, when $\mathbf{W}$ is used to describe a facial expression as

$$\mathbf{E}\boldsymbol{\varepsilon} \approx \mathbf{W}\mathbf{z} = z_1\mathbf{w}_1 + z_2\mathbf{w}_2 + \cdots + z_K\mathbf{w}_K, \tag{3}$$

only a subset of the coefficients $\{z_i\}_i$ are expected to be non-zero. While some of the expressions encoded in the components $\mathbf{w}_i$ can resemble AUs, in general, the coding system must be different from the FACS; otherwise, the non-additive AU combinations could not be reconstructed via a linear sum as in (3). Indeed, experiments show that some non-additive AU combinations receive a dedicated component (Section IV-C.1), although this does not imply that all non-additive combinations will receive a dedicated component. The exact nature of the components is determined by the procedure –the algorithms and the data– that is used while learning the coding system.

C. The Procedure for Learning the Localized Face Basis

We construct the model $\mathbf{W}$ from a dataset of N images with facial expression variations, in an unsupervised manner. As a first step, we perform 3DMM fitting on every image, and obtain the corresponding vectors of expression coefficients $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_N$. Then, we learn the model $\mathbf{W}$ by fitting a sparse dictionary [5], which amounts to solving an optimization problem where the objective is to minimize the discrepancy between the two representations of the same expressions –the representation via the 3DMM's expression model $\mathbf{E}$ and the targeted local basis $\mathbf{W}$–,

$$\sum_n \|\mathbf{E}\boldsymbol{\varepsilon}_n - \mathbf{W}\mathbf{z}_n\|_2 + \lambda \sum_n \|\mathbf{z}_n\|_1, \tag{4}$$

w.r.t. $\mathbf{W}$ and $\{\mathbf{z}_n\}$. During optimization, we also enforce the ℓ_2 norm of the components $\mathbf{w}_i$ to not exceed 1, otherwise they can grow unboundedly and compromise the desired sparsity pattern by allowing very small but non-zero values in $\mathbf{z}_n$ to have a significant impact [5].

The ℓ_1 penalty $\lambda \sum_n \|\mathbf{z}_n\|_1$ in (4) ensures that the representation via $\mathbf{W}$ will have a sparse activation pattern [5], satisfying one of the two properties of FACS that we try to emulate (Section III-B). The other property that we are after, *i.e.*, having localized expression units, is not guaranteed by the terms in (4). One way to achieve this property is adding an ℓ_1 penalty on the expression units $\mathbf{w}_i$. While this practice is known to lead to spatially localized components [22], [42], [43] , there is no guarantee that all components will be localized—the resulting components may span the entire input [22]. The existence of such global expression units is against a key property of FACS that we try to emulate. Thus, we enforce a constraint that results in a set of expression units where every unit is localized. This constraint is based on using the landmarks corresponding to facial features, as explained below.

3DMMs typically contain L indices that correspond to a subset of the P mesh points that enclose the main facial features, namely the eyes, the brows, the nose and the mouth. For example, the iBUG-51 template [40] contains L=51 landmarks. We use these landmarks to ensure that every expression unit is localized. This is achieved by allocating each expression unit $\mathbf{w}_i$ to a specific facial feature a priori, and allowing it to contain movement along the landmark points of the corresponding facial feature but not the other landmarks. For example, suppose that the left brow is represented with L_{LB} landmark points. Then, each component $\mathbf{w}_i$ contains $3L_{\text{LB}}$ entries (*i.e.*, the 3D coordinates of L_{LB} points) that correspond to the left brow landmarks. If a specific $\mathbf{w}_i$ is allocated to the left brow, then the corresponding $3L_{\text{LB}}$

entries are allowed to be non-zero, but the entries corresponding to the other $3(L-L_{\text{LB}})$ landmarks are forced to be zero. In this way, all the K expression units are divided into six groups a priori, corresponding to the six facial features (two brows, two eyes, nose, mouth). Experiments show that the components learned in this way contain expression units that are localized—they control predominantly one facial feature (Section IV-C.1).

IV. Experiments

We first investigate the learnt Facial BUs and compare them to FACS AUs. Next, we show the utility of the proposed coding system in a clinical application by reporting results of classification experiments (AUT vs. NT) from conversational videos of adolescents. We provide comparisons with the most widely used automated FACS coder, namely, OpenFace [7].

A. Datasets

Dictionary Learning. We learned the Facial Basis by using the CK+ [44] and MMI [45] datasets. While spontaneous datasets such as DISFA [46] or BP4D+ [17] can also be used, one must remove the video parts without facial movements and stratify the facial expressions before doing so, to ensure that the underlying reconstruction algorithm is not biased by the video parts without expressions and that expressions with low as well as high base rate are covered. The CK+ and MMI datasets are readily cropped to facial movements and contain a large variety of expressions, eliminating the need for preprocessing steps that can affect the learned expression components.

Clinical Application. We conducted clinical classification experiments on two datasets to assess the reproducibility of our findings in two different contexts, namely in-person face-to-face (F2F) conversations and remote room-to-room (R2R) conversations. The second context included videos collected through a lagless, cable-connected setup mimicking online video conferencing with ideal connectivity conditions between two separate rooms. English speaking participants included adolescents (age 12-17) drawn from a larger sample who participated in studies at The Children's Hospital of Philadelphia. The data collection procedure was approved by the Institutional Review Board (IRB) of The Children's Hospital of Philadelphia. The F2F dataset included 42 participants (AUT n = 21 [13 male]; NT n = 21 [13 male]), and the R2R dataset included 97 participants (AUT n = 49 [30 male]; NT n = 48 [29 male]). For both datasets, groups were matched on age, sex ratio, and intelligence quotient (IQ) (Table I). Autism diagnoses were confirmed through the best clinical judgment of a licensed psychologist using all available information, including administration of the Autism Diagnostic Observation Schedule (ADOS-2) [28]. NT participants did not have any history of mental health diagnosis per self- or parent-report.

Participants completed the Contextual Assessment of Social Skills [47], a brief 3-4 minute semi-structured "get-to-know-you" conversation with a member of the research staff (confederates). Confederates were research assistants or students from the lab, assigned based on availability, whom the participant had not previously met. Participants and confederates were seated across from each other with two video cameras placed in between to record synchronized frontal videos of each person at 30 frames per second. Confederates were instructed to appear interested and engaged but not carry the conversation (*i.e.*, speak no more than 50% of the time and wait 5 seconds to re-initiate the conversation after a lapse).

TABLE I

The (mean) age and IQ; and number of female and male participants in the two study samples. p values indicate possible group differences.

	Face-to-Face (F2F)			Room-to-room (R2R)		
	AUT	NT	p val.	AUT	NT	p val.
Age	14.3	14.2	0.90	15.2	15.1	0.78
IQ	98.1	108.0	0.06	113.0	112.0	0.78
F/M participants	8/13	8/13	1.00	19/30	19/29	1.00

B. Experimental Setup

Compared coding systems. We compare the prediction accuracy of four facial expression coding systems: FACS coding via OpenFace, Facial Basis (Section III-B), PCA (Section III-A) and local PCA. The latter contains feature-specific expression components like the BUs, but the components are still learned using PCA instead of sparse coding. For each coding system, we investigate how performance varies with the number of components used. Specifically, we report classification accuracy obtained using the first k components, where k is increased with increments of five (*i.e.*, $k = 5, 10, \ldots, K_{\text{exp}}$). For these experiments, the expression components are ordered according to their magnitude of activation on the datasets used for the experiments. That is, the kth expression component of a coding system is the one that had the kth highest magnitude across the F2F and R2R datasets.

Classification pipeline. Given the restrictions that stem from the clinical sample sizes, we conducted experiments using shallow classification pipelines. Also, to reliably tune the pipelines through (nested) cross validation, we aimed to minimize the number of hyperparameters and we used a linear SVM classifier, which contains only one parameter to be tuned, namely, the C parameter. The raw input data are defined by K facial expression signals and 3 head movement signals (*i.e.*, rotation angles). The head movements are included in FACS as facial action descriptors [1] and are known to carry meaningful communication cues that vary with mental health conditions [48], [49]. The facial expression signals at every frame are obtained either using a data-driven coding system or OpenFace. The final input features are generated from these $Q = K + 3$ signals using (intra-person) windowed cross-correlation (WCC). WCC is a widely used approach in behavioral research [50] and can capture behavioral differences that are expected to exist between AUT and NT groups, such as typicality/atypicality of expressions [51] or level of integrating multiple components

of behavior [52]. WCC represents the behavior within a time window of T_w seconds through a vector of Q^2 features (all possible pairs of signals). The behavior over the entire video is represented as a Q^2-dimensional vector obtained by averaging over the feature vectors across all time windows within a video [31]. These Q^2-dimensional feature vectors are used with the SVM classifier. Results are reported with leave-one-out cross-validation, and the C parameter at each cross-validation fold is tuned with a 5-fold inner cross-validation (*i.e.*, nested cross validation).

3DMM fitting. The 3DMM fitting procedure needed for the Facial Basis (Section III-A) was devised to be computationally efficient while also representing expressions comprehensively. Existing deep learning-based 3DMM fitting software are usually computationally efficient but represent expressions with reduced capacity. For example, the methods based on the Basel Face Model (BFM) [53] topology may use only 29 [54], [55] of the $M = 79$ components of associated expression model [25]. Optimization-based methods such as 3DI [24], on the other hand, can use all expression components as they do not commit to a specific version of a 3DMM, but are prohibitively slow or require GPU during inference. Thus, we trained a ResNet, which predicts identity, pose and expression parameters (Section III-A) according to the BFM model from an image frame in a supervised fashion, where the labels are obtained by using the 3DI method. Specifically, we trained using a large dataset (*i.e.*, YouTube Faces [56] and CelebA [57] datasets combined) to minimize the L2 loss between the labels predicted by the ResNet and the labels obtained by 3DI prior to training.

Implementation details. The Facial Basis was learned using `scikit-learn`, which implements the minimization of (4) with its dictionary learning module. We set the λ coefficient of (4) to 0.2, and the number of components of the coding system $\mathbf{W}$ to K=50, as qualitative inspection suggested that increasing K further led to basis components that are highly similar. Each expression component in $\{\mathbf{w}_i\}_{i=1}^{K}$ corresponds to a specific facial feature, and for convenience, we associated a two-letter code to each of the components according the feature they correspond to (LB: left brow, RB: right brow, LE: left eye, RE: right eye, NO: nose and MO: mouth). Thus, each component $\mathbf{w}_i$ has a unique name, constructed by using the two-letter code and a number. For example, LE-3 is the third expression component associated with the left eye.

C. Results

1) The Learned Expression Components: Fig. 2 shows the expressions encoded by some of the FBUs, and suggests that many components correspond to plausible and interpretable facial movements—movements that are localized and can be generated by the human face. This is in contrast with the PCA-based expression basis of the 3DMMs, which contain expression components that are not localized and can be physically implausible (Fig. 1). The BUs show some visible similarities to FACS AUs, but they also diverge from them.

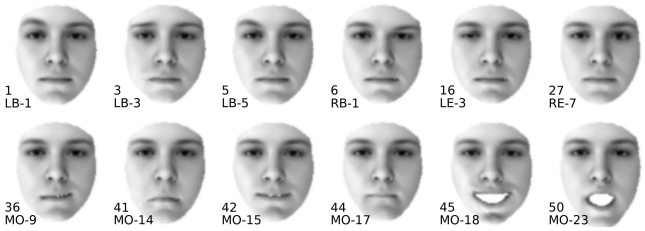

Fig. 2. The expression encoded by some Facial Basis Units (BUs).

TABLE II

LEFT: CORRESPONDENCES BETWEEN FACIAL BASIS UNITS (BUS) AND AUS. RIGHT: DESCRIPTION OF SPECIFIED AUS. FOLLOWING FACS ANNOTATION, ASYMMETRIC AUS ARE DENOTED BY 'L' (IF ONLY ON THE LEFT HEMIFACE) OR 'R' (IF ONLY ON THE RIGHT HEMIFACE).

Facial Basis Unit (BU)	AU Combination
LB-1	AU L1+L2
LB-3	AU L1+L4
LB-5	AU L2
RB-1	AU R4
LE-3	AU 41
RE-7	AU 5
MO-9	AU R12+R14A
MO-14	AU 15
MO-15	AU 12
MO-17	AU 24
MO-18	AU 12+27
MO-23	AU 26

AU Code	Description
AU 1	Inner Brow Raiser
AU 2	Outer Brow Raiser
AU 4	Brow Lowerer
AU 5	Upper Lid Raiser
AU 12	Lip Corner Puller
AU 14	Dimpler
AU 15	Lip Corner Depressor
AU 24	Lip Pressor
AU 26	Jaw Drop
AU 27	Mouth Stretch
AU 41	Lid Droop

The existence of differences between these two coding systems is expected by design, since our goal is not to replicate FACS but create an alternative coding system that can represent expressions as a linear sum, which cannot be done by FACS due to the non-additivity of some AUs (Section III-B). For example, one of the Facial BUs resembles AU 1+4 (LB-3 in Fig. 2), which is a reasonable outcome, as AU 1 and AU 4 are not additive, and a linear model that contains only these units would not be able to represent their combination. Table II shows more examples of similar-looking BUs and AUs (or AU combinations). Fig. 3 shows that despite the lack of one-to-one matching between FACS AUs and the Facial BUs, both systems can be used to infer localized movements in facial videos. The visualization of all 50 BUs is provided at `github.com/sariyanidi/FacialBasis`.

A critical advantage of the Facial Basis compared to automated FACS software is that it readily supports the analysis of asymmetric expressions (see modifiers R and L in Table II). Automated FACS software outputs typically do not provide predictions for AUs that occur only on one hemiface [7], [16], [19], in part due to the difficulties related to eliciting asymmetric facial expressions. Being able to capture asymmetrical expressions or quantifying the degree of symmetry can be important for predicting autism [58] or studying neurological conditions such as cerebral palsy.

2) Clinical Classification Results: Fig. 4 reports the AUT vs. NT classification performance of the compared coding systems, and shows how performance varies with the number of expression components used. On the F2F sample, the

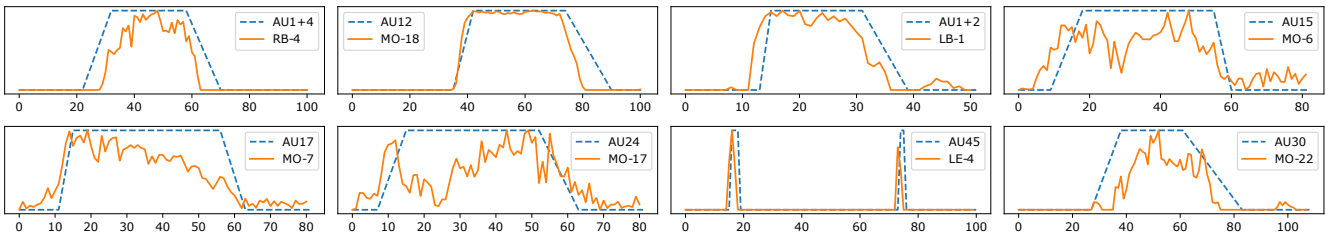

Fig. 3. The AU labels (ground truth) from videos of the MMI dataset and the BU coefficients of the corresponding expression units, plotted over time. Results suggest that both the FACS AUs and the Facial BUs can be used to infer localized movements.

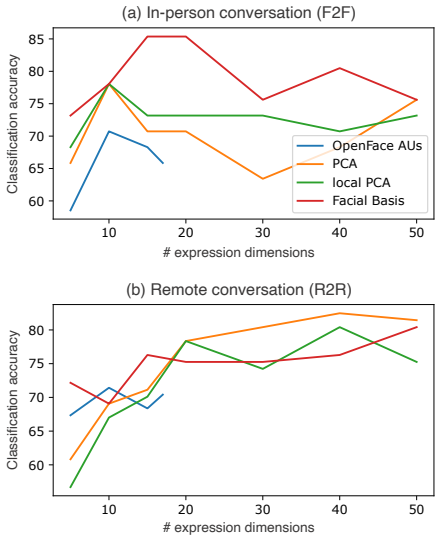

Fig. 4. AUT vs. NT classification results of the compared coding systems w.r.t. the number of expression components used per coding system.

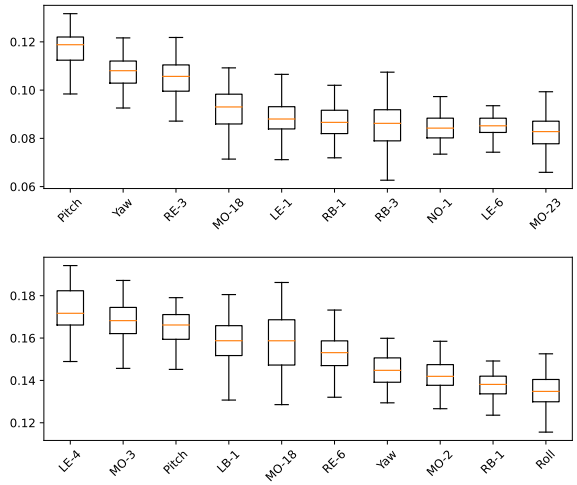

Fig. 5. Average feature weights of behavioral components that include head movement (Pitch, Yaw, Roll angles) as well as Facial BUs. Top: in-person sample (F2F); bottom: remote (R2R) sample.

highest classification accuracy is achieved by the Facial Basis (Fig. 4a). Of note, all data-driven coding systems outperform OpenFace. All methods reach their peak performance with 10-15 expression components on the F2F sample, suggesting that a relatively small number of expression units may suffice to achieve peak accuracy. However, the trend is different for the R2R sample. All methods other than automated FACS reach their peak performance around 40-50 units (Fig. 4b). The number of components examined for FACS was limited to 17, the maximum number of AU intensity estimates provided by OpenFace.

The different trends between F2F (Fig. 4a) and R2R (Fig. 4b) suggest that behavioral characteristics of in-person communication may be different from those of a remote, computer-based communication, even when the latter is lagless. To further investigate potential differences caused by the communication medium, we performed classification experiments using only head movements. On the F2F sample, the classification accuracy of head movements alone was 73%, which is consistent with previous literature suggesting that head movements contain rich information to classify between the AUT and NT groups [49], [59], [60]. However, the classification accuracy on the R2R sample was only 32%, providing further evidence about the existence of behavioral differences between in-person vs. remote conversations.

Fig. 5 provides further insights about the differences between the two recording setups by illustrating the ten Facial BUs that received the highest weight by the SVM classifiers. Since feature weights depend on the training sample, we plot a distribution (*i.e.*, boxplot) per component, constructed by using the feature weights of the corresponding component across 100 randomly picked subsamples of the training data. The head movement components (pitch and yaw rotation) hold the highest weight in the F2F sample but not in the R2R sample, further supporting the importance of head movements during in-person communication. The next most important feature in the F2F sample is RE-3, which corresponds to closing eyelids (see Supplementary Video). The high weight of this component may be attributed to atypical eye-blink patterns that have been observed in autism [61], [62] or its indirect link to social gaze behavior [63]. The feature with the next highest weight is MO-18, which corresponds to a smile with open mouth (Fig. 2), and the importance of this feature may be explained by the atypicality of smile [64] in youth with autism or reduced inter-personal affect coordination [32] in autism.

An important detail in Fig. 5 is that the classifier weights for the R2R setup show slower decay, further supporting the

need of a richer representation with an increased number of components. While in general the features that lead to the highest classification accuracy depend on the recording setup, MO-18 is one of the components that is in the top five of both recording setups, and another one is the pitch rotation, which occurs with head nodding. Another high-ranking feature in the R2R sample is MO-3, which includes a lip corner movement that can also be observed with a smile. It must be noted that the weight corresponding to each expression component in Fig. 5 is depicted after averaging all the cross-correlation features that include the component. Therefore, the listed expression components are not necessarily important in isolation, but in combination with other components, since the cross-correlation features encode the relationship between different behavioral components (Section IV-B). A more nuanced analysis would require the inspection of the individual features without averaging. However, such an analysis requires larger samples, as the weight of individual features had too much variance and instability in our sample, which is expected when the number of individual features is high [65], as in our pipeline ($Q^2{=}53^2{=}2809$).

In sum, our experiments provide three critical insights. First, data-driven coding outperforms the most widely used AU detector in predicting autism. Second, automated facial expression analysis in mental health may require or benefit from more expression coefficients than those provided by automated FACS coders. Third, in-person and remote communication likely have different behavioral dynamics.

V. Limitations

The presented coding system suffers from four limitations that reduce its interpretability or representation power. First, while most Facial BUs are interpretable and physically plausible, there are a number of units that are difficult to semantically describe or unlikely to be generated by a face (*e.g.*, MO-4, MO-9 and MO-12 in supplementary video). These units can be reduced or eliminated by augmenting the objective function with loss terms that enforce expression units to be physically plausible by imposing certain anatomical constraints [66]. The second limitation is that some of the expression units describe very similar movements, and the lack of distinct differences between expression units can complicate analyses (Section II-B). This issue can be remedied by augmenting the objective function (4) with terms to prevent BUs from being highly similar, or by creating a hierarchy, akin to hierarchical clustering [67], where expression units with a high degree of similarity are clustered together. Third, our classification experiments are based on a spontaneous task, but the Facial Basis is currently trained with only posed expressions. To better represent naturalistic expressions, datasets collected from spontaneous tasks or interactions [17], [46] can be included. The inclusion of spontaneous videos may require additional pre-processing procedures or using alternative loss function (*e.g.*, wing loss [68]) to ensure that expressions with low as well as high base rates are represented by the learned coding system. Finally, the Facial BUs are limited in their ability to capture wrinkles and furrows that may occur with expressions. This is likely caused by the fact that the 3DMM that we use to model the 3D face, BFM, has a relatively low resolution, as it represents the facial region with approximately 20k points. This limitation may be remedied by the usage of 3DMMs with higher resolution, such as the FaceScape [69], which represents 3D face with approximately 2 million points.

A limitation on the interpretation of the differences between the R2R and F2F setups is that recordings were drawn from two independent samples. While these samples are demographically and clinically similar, it is possible that underlying differences in the participants in each sample drove differences, rather than recording setup.

VI. Discussion and Outlook

Our results suggest that a data-driven coding system provides an interpretable approach for encoding facial behavior, and outperforms a standard FACS coder in the scope of predicting autism. Thus, we show that a promising new paradigm is to develop alternative coding systems that mimic FACS rather than developing automated coders that mimic FACS experts, particularly because the accuracy of the latter approach may be saturating after more than forty years of research.

An immediate impact of the proposed study is to use the specific coding system, namely Facial Basis, for studying mental conditions and other behavioral research questions (*e.g.*, emotions). It should be noted, however, that the proposed system is simply a starting point and its current limitations (Section V) suggest that its output must be interpreted with caution. The restricted dataset used during its training may not be representative across a variety of research questions (*e.g.*, detection of pain or prediction of emotional states).

The possibly more significant but longer-term impact of our study is the construction of a universal coding system similar to FACS that can serve as a common language for quantifying behavior across studies, contexts and mental health conditions. A system of this kind would encode facial expressions comprehensively, reliably, and validly, while also proving successful across a variety of mental health applications. As such, the construction of a universal coding system must be the result of more intense efforts spanning multiple studies in diverse populations. The dataset that is used for such a coding system must be large enough to cover virtually any facial action that can be generated by a face, while the learning procedure must ensure that expressions with low as well as high base rate are represented.

Ethical Impact Statement

The main objective of this study is to advance behavioral and mental health research by providing scientists with a tool that encodes facial behavior more comprehensively than existing tools. This tool, however, could be used by malevolent actors who want to improve skills and capabilities that are detrimental to society, such as deception with facial behavior.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[2] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 04 2005. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

[3] "Facial Action Coding System - Paul Ekman Group," https://www.paulekman.com/facial-action-coding-system/, accessed: 2024-12-19.

[4] I. O. Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji, "Crossing domains for au coding: Perspectives, approaches, and measures," *IEEE transactions on biometrics, behavior, and identity science*, vol. 2, no. 2, pp. 158–171, 2020.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding." *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.

[6] Y. L. de Nocker and C. K. Toolan, "Using telehealth to provide interventions for children with asd: A systematic review," *Review Journal of Autism and Developmental Disorders*, vol. 10, no. 1, pp. 82–112, 2023.

[7] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10. [Online]. Available: https://ieeexplore.ieee.org/document/7477553

[8] J. J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Automated facial expression recognition based on facs action units," in *Proceedings third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 390–395.

[9] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[10] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, no. 2, pp. 253–263, 1999.

[11] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[12] Y. Chang and S. Wang, "Knowledge-Driven Self-Supervised Representation Learning for Facial Action Unit Recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20 385–20 394, iSSN: 2575-7075. [Online]. Available: https://ieeexplore.ieee.org/document/9880051

[13] Y. Yin, D. Chang, G. Song, S. Sang, T. Zhi, J. Liu, L. Luo, and M. Soleymani, "FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2024, pp. 6087–6096. [Online]. Available: https://ieeexplore.ieee.org/document/10483673/

[14] B. Ma, R. An, W. Zhang, Y. Ding, Z. Zhao, R. Zhang, T. Lv, C. Fan, and Z. Hu, "Facial Action Unit Detection and Intensity Estimation From Self-Supervised Representation," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1669–1683, Jul. 2024, conference Name: IEEE Transactions on Affective Computing. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10439628

[15] X. Liu, K. Yuan, X. Niu, J. Shi, Z. Yu, H. Yue, and J. Yang, "Multi-scale Promoted Self-adjusting Correlation Learning for Facial Action Unit Detection," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024, conference Name: IEEE Transactions on Affective Computing. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10679925

[16] S. Hinduja, I. O. Ertugrul, M. Bilalpur, D. S. Messinger, and J. F. Cohn, "PyAFAR: Python-based Automated Facial Action Recognition library for use in Infants and Adults." IEEE Computer Society, Sep. 2023, pp. 1–3. [Online]. Available: https://www.computer.org/csdl/proceedings-article/aciiw/2023/10388108/1TKQZrMxwzK

[17] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.

[18] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 881–888.

[19] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox," *Affective Science*, vol. 4, no. 4, pp. 781–796, 2023.

[20] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system, a human face," *ETC What is the ETC*, 2002.

[21] J. F. Cohn, "Foundations of human computing: Facial expression and emotion," in *Proceedings of the 8th international conference on Multimodal interfaces*, 2006, pp. 233–238.

[22] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1965–1978, 2017.

[23] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani *et al.*, "3d morphable face models—past, present, and future," vol. 39, no. 5, pp. 1–38, 2020.

[24] E. Sariyanidi, C. J. Zampella, R. T. Schultz, and B. Tunç, "Inequality-constrained 3d morphable face model fitting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[25] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," vol. 20, no. 3, pp. 413–425, 2013.

[26] K. Anjyo, *Blendshape Facial Animation*. Cham: Springer International Publishing, 2018, pp. 2145–2155. [Online]. Available: https://doi.org/10.1007/978-3-319-14418-4_2

[27] D. Curran-Everett, "Multiple comparisons: philosophies and illustrations," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 279, no. 1, pp. R1–R8, 2000.

[28] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, *Autism Diagnostic Observation Schedule, Second Edition: ADOS-2*, 2012.

[29] J. C. Koehler, M. S. Dong, A. M. Bierlich, S. Fischer, J. Späth, I. S. Plank, N. Koutsouleris, and C. M. Falter-Wagner, "Machine learning classification of autism spectrum disorder based on reciprocity in naturalistic social interactions," *Translational Psychiatry*, vol. 14, no. 1, p. 76, 2024.

[30] N. Zhang, M. Ruan, S. Wang, L. Paul, and X. Li, "Discriminative few shot learning of facial dynamics in interview videos for autism trait classification," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1110–1124, 2022.

[31] E. Sariyanidi, C. J. Zampella, E. DeJardin, J. D. Herrington, R. T. Schultz, and B. Tunc, "Comparison of human experts and ai in predicting autism from facial behavior," in *CEUR workshop proceedings*, vol. 3359, no. ITAH. NIH Public Access, 2023, p. 48.

[32] C. J. Zampella, L. Bennetto, and J. D. Herrington, "Computer vision analysis of reduced interpersonal affect coordination in youth with autism spectrum disorder," *Autism Research*, vol. 13, no. 12, pp. 2133–2142, 2020.

[33] Z. Zhao, Z. Zhu, X. Zhang, H. Tang, J. Xing, X. Hu, J. Lu, Q. Peng, and X. Qu, "Atypical head movement during face-to-face interaction in children with autism spectrum disorder," *Autism Research*, vol. 14, no. 6, pp. 1197–1208, 2021.

[34] H. Drimalla, N. Landwehr, I. Baskow, B. Behnia, S. Roepke, I. Dziobek, and T. Scheffer, "Detecting autism by analyzing a simulated social interaction," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 2019, pp. 193–208.

[35] B. Turan, P. Algedik Demirayak, E. Yildirim Demirdogen, M. Gulsen, H. C. Cubukcu, M. Guler, H. Alarslan, A. E. Yilmaz, and O. B. Dursun, "Toward the detection of reduced emotion expression intensity: an autism sibling study," *Journal of Clinical and Experimental Neuropsychology*, vol. 45, no. 3, pp. 219–229, 2023.

[36] N. I. Solórzano Alcívar, D. F. Paillacho Chiluiza, M. X. Arce Sierra, A. J. Pincay Lino, and E. A. Eras Zamora, "Facial expression analysis in children with autism spectrum disorder using a refined human-robot-game platform for active learning," *Behaviour & Information Technology*, pp. 1–13, 2024.

[37] Y. Li, H. Liu, H. Feng, X. Shen, Z. Chen, W. Luo, N. Li, and S. Tan, "Capturing fear through eyes to aid in restoring social functioning for

neuropsychiatric disorders: Machine learning research evidence from the emotion face database," 2024.

[38] W. Saakyan, M. Norden, L. Herrmann, S. Kirsch, M. Lin, S. Guendelman, I. Dziobek, and H. Drimalla, "On scalable and interpretable autism detection from social interaction behavior," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[39] G. Alvari, C. Furlanello, and P. Venuti, "Is smiling the key? machine learning analytics detect subtle patterns in micro-expressions of infants with asd," *Journal of clinical medicine*, vol. 10, no. 8, p. 1776, 2021.

[40] E. Sariyanidi, C. J. Zampella, R. T. Schultz, and B. Tunc, "Can facial pose and expression be separated with weak perspective camera?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7173–7182.

[41] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[42] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[43] T. Kim, G. Shakhnarovich, and R. Urtasun, "Sparse coding for learning interpretable spatio-temporal primitives," *Advances in neural information processing systems*, vol. 23, 2010.

[44] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[45] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, pp. 5–pp.

[46] 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*, vol. 4, no. 2, pp. 151–160, 2013.

[47] A. B. Ratto, L. Turner-Brown, B. M. Rupp, G. B. Mesibov, and D. L. Penn, "Development of the contextual assessment of social skills (cass): A role play measure of social skill for individuals with high-functioning autism," *Journal of Autism and Developmental Disorders*, vol. 41, pp. 1277–1286, 2011.

[48] T. Horigome, B. Sumali, M. Kitazawa, M. Yoshimura, K.-c. Liang, Y. Tazawa, T. Fujita, M. Mimura, and T. Kishimoto, "Evaluating the severity of depressive symptoms using upper body motion captured by rgb-depth sensors and machine learning in a clinical interview setting: a preliminary study," *Comprehensive psychiatry*, vol. 98, p. 152169, 2020.

[49] D. Q. McDonald, E. Sariyanidi, C. J. Zampella, E. Dejardin, J. D. Herrington, R. T. Schultz, and B. Tunc, "Predicting autism from head movement patterns during naturalistic social interactions," in *Proceedings of the 2023 7th International Conference on Medical and Health Informatics*, 2023, pp. 55–60.

[50] S. M. Boker, J. L. Rotondo, M. Xu, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series." *Psychological methods*, vol. 7, no. 3, p. 338, 2002.

[51] R. Brewer, F. Biotti, C. Catmur, C. Press, F. Happé, R. Cook, and G. Bird, "Can neurotypical individuals read autistic facial expressions? atypical production of emotional facial expressions in autism spectrum disorders," *Autism Research*, vol. 9, no. 2, pp. 262–271, 2016.

[52] V. Hus and C. Lord, "The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores," *Journal of autism and developmental disorders*, vol. 44, pp. 1996–2012, 2014.

[53] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

[54] F.-J. Chang, A. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in *13th IEEE Conference on Automatic Face and Gesture Recognition*, 2018.

[55] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *European Conf. on Computer Vision (ECCV)*, 2020.

[56] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*. IEEE, 2011, pp. 529–534.

[57] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.

[58] T. Guha, Z. Yang, A. Ramakrishna, R. B. Grossman, D. Hedley, S. Lee, and S. S. Narayanan, "On quantifying facial expression-related atypicality of children with autism spectrum disorder," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 803–807.

[59] K. B. Martin, Z. Hammal, G. Ren, J. F. Cohn, J. Cassell, M. Ogihara, J. C. Britton, A. Gutierrez, and D. S. Messinger, "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Molecular autism*, vol. 9, pp. 1–10, 2018.

[60] M. Gokmen, E. Sariyanidi, L. Yankowitz, C. J. Zampella, R. T. Schultz, and B. Tunç, "Detecting autism from head movements using kinesics," in *Proceedings of the 26th International Conference on Multimodal Interaction*, 2024, pp. 350–354.

[61] L. L. Sears, P. R. Finn, and J. E. Steinmetz, "Abnormal classical eye-blink conditioning in autism," *Journal of autism and developmental disorders*, vol. 24, no. 6, pp. 737–751, 1994.

[62] P. R. Krishnappa Babu, V. Aikat, J. M. Di Martino, Z. Chang, S. Perochon, S. Espinosa, R. Aiello, K. LH Carpenter, S. Compton, N. Davis *et al.*, "Blink rate and facial orientation reveal distinctive patterns of attentional engagement in autistic toddlers: a digital phenotyping approach," *Scientific Reports*, vol. 13, no. 1, p. 7158, 2023.

[63] L. Yankowitz, M. K. Pargi, E. Dejardin, C. J. Zampella, W. Guthrie, J. Pandey, J. Bartley, G. K, D. Chen, D. Q. McDonald, A. Manakiwala, and et al., "Computational measurement of social gaze during naturalistic conversations in autism," Dec 2024. [Online]. Available: osf.io/preprints/psyarxiv/rjvmx

[64] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," in *2013 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2013, pp. 1–6.

[65] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[66] A.-E. Ichim, P. Kadleček, L. Kavan, and M. Pauly, "Phace: Physics-based face modeling and animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[67] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[68] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2235–2245.

[69] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 601–610.