Published in final edited form as:
CEUR Workshop Proc. 2023 March ; 3359(ITAH): 48–57.

Comparison of Human Experts and AI in Predicting Autism from Facial Behavior

Evangelos Sariyanidi[1], **Casey J. Zampella**[1], **Ellis DeJardin**[1], **John D. Herrington**[1,2], **Robert T. Schultz**[1,2], **Birkan Tunc**[1,2]

[1]Center for Autism Research, The Children's Hospital of Philadelphia, United States

[2]University of Pennsylvania, United States

Abstract

Advances in computational behavior analysis via artificial intelligence (AI) promise to improve mental healthcare services by providing clinicians with tools to assist diagnosis or measurement of treatment outcomes. This potential has spurred an increasing number of studies in which automated pipelines predict diagnoses of mental health conditions. However, a fundamental question remains unanswered: How do the predictions of the AI algorithms correspond and compare with the predictions of humans? This is a critical question if AI technology is to be used as an assistive tool, because the utility of an AI algorithm would be negligible if it provides little information beyond what clinicians can readily infer. In this paper, we compare the performance of 19 human raters (8 autism experts and 11 non-experts) and that of an AI algorithm in terms of predicting autism diagnosis from short (3-minute) videos of $N = 42$ participants in a naturalistic conversation. Results show that the AI algorithm achieves an average accuracy of 80.5%, which is comparable to that of clinicians with expertise in autism (83.1%) and clinical research staff without specialized expertise (78.3%). Critically, diagnoses that were inaccurately predicted by most humans (experts and non-experts, alike) were typically correctly predicted by AI. Our results highlight the potential of AI as an assistive tool that can augment clinician diagnostic decision-making.

Keywords

autism; assistive healthcare technologies; digital phenotyping

1. Introduction

Modern medical disciplines typically rely on a variety of technological tools to assist in diagnosis and monitor treatment progress. From brain imaging technologies to blood and genetic tests, instruments that assist medical decision-makers are a cornerstone of modern medicine. In the domain of psychiatry and psychology, however, medical decision-making relies nearly exclusively on observational or paper-and-pencil instruments. Thus, recent

sariyanide@chop.edu (E. Sariyanidi); zampellac@chop.edu (C. J. Zampella); dejardine@chop.edu (E. DeJardin); herringtonj@chop.edu (J. D. Herrington); schultzrt@chop.edu (R. T. Schultz); tuncb@chop.edu (B. Tunc).

advances in computer vision and artificial intelligence (AI) are poised to rapidly advance research and clinical decision-making in psychiatry by introducing reliable and granular tools within a new paradigm: computational behavior analysis [1, 2, 3, 4, 5]. Such tools can capture and quantify human behavior with extraordinary precision, even from brief video recordings.

Autism spectrum disorder (ASD), like nearly all psychiatric conditions, is defined by observable behavioral cues—what a person does well or not well, too little or too much. Its core traits include observable differences in social communication, social reciprocity, nonverbal communication, and relationships, as well as restricted patterns of interests and activities [6]. The current reliance on assessment and interpretation of overt behavior makes autism an excellent candidate for computational behavior analysis approaches. Coupling computationally-derived biomarkers with expert clinician judgment may provide an extremely potent approach to autism care, by enhancing the currently limited reliability of clinical assessments (*e.g.,* DSM-5 field trials Kappa = 0.69) [7], shortening lengthy diagnostic evaluations, and improving sensitivity for capturing change over the course of treatment and development.

This potential has spurred a plethora of studies that aim to diagnose autism via AI pipelines based on various behavioral modalities and sensors [8]. Notably, to our knowledge, no study has directly compared AI algorithms and human raters with respect to overall predictive capacity or specific decisions on individual cases. A comparison of this kind is important when it comes to using AI as an assistive technology for clinical decision-making, as it can determine whether or not AI provides significant incremental utility beyond existing tools. AI algorithms can maximize and cooperate synergistically with human assessment by complementing and augmenting human decisions. On the other hand, clinicians would have little interest in or benefit from incorporating AI algorithms if their decisions –and errors– highly overlap with their own. We aim to address this issue by examining whether or not AI detects diagnostic indicators that may go unnoticed by human observation.

In this paper, our main contribution is comparing the performance of AI and humans with knowledge of autism in accurately classifying autism from a 3-minute get-to-know-you conversation with a non-clinician conversation partner. Specifically, we implemented a computer vision pipeline for predicting autism using features of facial behavior during conversations with a sample of $N = 42$ adults − 15 individuals with autism spectrum disorder (ASD) and 27 neurotypical (NT) individuals. We then recruited a total of 19 human raters (8 expert clinicians, 11 non-experts with experience with autism) to predict the diagnostic status of the same participants. The expert raters were doctoral level clinicians with extensive training on autism, while most of the non-experts were BA level researchers still learning about autism. Raters watched the same videos of participants' faces during conversations that were fed to the computer vision pipeline, without sound to allow for a fairer comparison with the AI algorithm.

Results suggest that the AI pipeline based on participant facial behavior predicts diagnostic status with 80.5% accuracy. This accuracy was comparable to the 80.3% overall accuracy achieved by human raters (83.1% for experts and 78.3% for non-experts), demonstrating the potential of AI to detect facial behavioral patterns that differentiate adults with autism from neurotypical peers in the context of a casual, get-to-know-you conversation. Moreover, we show that the prediction errors of AI and humans had little overlap, indicating that the AI can provide complementary information that could prompt and assist clinicians with their evaluations and decision-making. The fact that all the results of this paper are extracted from a brief naturalistic conversation is a significant contribution, as a 3-minute conversation with a non-expert is a highly scalable paradigm, and thus a promising option as a screening or (preliminary) diagnostic procedure. The results of this paper motivate further research efforts to understand the decision mechanisms of AI algorithms, particularly for uncovering subtle behavioral patterns in psychiatric conditions.

2. Participants and Procedure

Forty-four adults participated in the present study (ASD: n=17, NT: n=27, all native and fluent English speakers). Participant groups did not differ significantly on mean chronological age, full-scale IQ estimates (WASI-II) [9], verbal IQ estimates, or sex ratio (Table 1). Participant diagnostic status (ASD or NT) was confirmed as part of this study using the Clinical Best Estimate process [10], informed by the Autism Diagnostic Observation Schedule - 2nd Edition, Module 4 (ADOS-2) [11] and adhering to DSM-V criteria for ASD [12]. All aspects of the study were approved by the Institutional Review Board The Children’s Hospital of Philadelphia (CHOP). Two participants were excluded from analysis due to their lack of consent for this particular set of experiments or their data being unavailable for processing, yielding a final sample of 42 participants (ASD: N=15, NT: N=27).

Participants underwent a battery of tasks that assessed social communication competence, including a slightly modified version of the Contextual Assessment of Social Skills (CASS) [13]. The CASS is a semi-structured assessment of conversational ability designed to mimic real-life first-time encounters. Participants engaged in two 3-minute face-to-face conversations with two different confederates (research staff, blind to participant diagnostic status and unaware of the dependent variables of interest). In the first conversation (interested condition), the confederate demonstrates social interest by engaging both verbally and non-verbally in the conversation. In the second conversation (bored condition), the confederate indicates boredom and disengagement both verbally (*e.g.,* one-word answers, limited follow-up questions) and physically (*e.g.,* neutral affect, limited eye-contact and gestures). All analyses throughout this paper are based on the interested condition only.

During the CASS, participants and confederates were seated facing one another. Audio and video of the CASS were recorded using an in-house device comprising two 1080p HD (30 fps) cameras (Fig. 1), which was placed between the participant and confederate on a floor stand. The two cameras of the device point in opposite directions to allow simultaneous recording of the participant and the confederate. However, the AI analyses in this paper are conducted on the video data of the participant only. In other words, even if the context of the conversation is *dyadic*, our AI-based analysis is not dyadic since it discards the information from the confederate and focuses only on the participant. We refer to this type of analysis as *monadic* analysis.

CASS confederates included 10 undergraduate students or BA-level research assistants (3 males, 7 females, all native English speakers). Confederates were semi-randomly selected, based on availability and clinical judgment. In order to provide opportunities for participants to initiate and develop the conversation, confederates were trained to speak for no more than 50% of the time and to wait 10s to initiate the conversation. If conversational pauses occurred, confederates were trained to wait 5s before re-initiating the conversation. Otherwise, confederates were told to simply naturally engage in the conversation. Prior to each conversation, study staff provided the following prompt to the participants and confederates before leaving the room: "Thank you both so much for coming in today. Right now, you will have 3 minutes to talk and get to know each other, and then I will come back into the room."

3. Prediction of Autism Diagnosis

3.1. Human Raters

We recruited a total of 19 human raters to view the videos from the sample of $N = 42$ participants. Eight of the raters were autism clinical experts, doctoral level clinicians with extensive training at the Center for Autism Research (CAR) of CHOP. The remaining 11 (non-expert) raters had some familiarity with autism but not specialized training and worked at CAR. Most of these non-expert raters were BA-level psychology students learning about autism.

The videos that were shown to the human raters were prepared as follows: First, we cropped the videos of the participant and their corresponding confederate conversation partner so that only the heads and necks were visible. Next, we combined the synchronized videos of the heads/faces of the participant and confederate into a single video file per participant such that participant and confederate were positioned side by side (Fig. 1, right). The audio was removed in order to allow human raters to focus on the facial behavior, as was the case for the AI algorithm. The videos for all $N = 42$ participants were presented to human raters in a random order on high resolution monitors.

Raters were instructed to watch each video just once and to make a decision as to whether the study participant had autism or not. They were told that all participants were either confirmed to have autism through clinical evaluation by a licensed expert, or were recruited specifically as neurotypical controls (*i.e.,* clear cases of individuals without autism). Raters were not allowed to go back and review earlier videos. They were instructed to watch all videos within 1 to 3 viewing sessions, with nearly all being completed in 1 or 2 sessions.

3.2. Computer vision

3.2.1. Quantification of facial behavior—Our goal is to quantify all observable facial behavior of a participant, which includes facial expressions and head movements. Also, we did not want to limit analysis to emotion-related expressions (*e.g.,* the six basic emotions), as other kinds of facial movements (*e.g.,* communicative expressions, speech-related mouth movements) are also important for diagnosing autism [14]. Therefore, we quantify behavior using a 3D morphable model (3DMM) [15] as 3DMMs contain expression bases (*e.g.,* [16]) that can quantify any facial movement. Moreover, 3DMMs can simultaneously model facial identity, pose, and expression. This increases the precision of parsing facial expressions and head movements, since the effect of identity (*i.e.,* identity bias [17]) is reduced when modeled and thus explained away. Specifically, we use the 3DI method [18], as it can learn identity from multiple frames and thus model and remove its effect more accurately. Moreover, 3DI can take the parameters of the camera as input, which is critical for increasing the accuracy with which facial expressions and pose are decoupled [19].

A 3DMM method produces a dense mesh of P three-dimensional points $\mathbf{X} \in \mathbb{R}^{3\times \mathrm{P}}$ to represent the face in a given video frame $\mathbf{I}$. (P is 23, 660 for the 3DI method). This 3D mesh is a function of the facial pose (*i.e.,* a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and a translation vector $\boldsymbol{\tau} \in \mathbb{R}^{3\times 1}$), the facial identity of the person $\overline{\mathbf{X}}$ and the facial expression variation in the image $\boldsymbol{\Delta}\mathbf{X} \in \mathbb{R}^{3\times P}$:

$$\mathbf{X} = \mathbf{R}\left(\overline{\mathbf{X}} + \boldsymbol{\Delta}\mathbf{X}\right) + \mathbf{T}, \tag{1}$$

where the columns of the matrix $\mathbf{T} \in \mathbb{R}^{3\times \mathrm{P}}$ are identically $\boldsymbol{\tau}$. The matrices of interest in the scope of our study are the matrix of head rotation $\mathbf{R}$ and the expression variation, $\boldsymbol{\Delta}\mathbf{X}$. 3DMMs represent expression variation as a linear sum, $\boldsymbol{\Delta}\mathbf{X} = \mathbf{W}\boldsymbol{\mathcal{E}}$, where $\mathcal{E} \in \mathbb{R}^{K\times 1}$ is the vector representing the expression. The expression basis $\mathbf{W}$ used by 3DI method is constructed via PCA [16], which limits the interpretability as PCA components are not localized–we cannot associate any PCA component with a specific facial region. To make the results of our study more interpretable, we modified the expression model in a way that the resultant expression model, $\mathbf{W}'$, contains 60 localized basis components as shown in Fig. 2. Using this model, we represent the expression variation in the image with the vector $\mathcal{E}'$ that minimizes the norm $//\boldsymbol{\Delta}\mathbf{X} - \mathbf{W}'\mathcal{E}'//_2$. We ignore the 7 components that correspond to the nose and cheek regions (Fig. 2), and we finally represent the expression variation in a video of T frames with a matrix $\mathbf{E}$ of size $T \times 53$, obtained by horizontally concatenating the expression vectors from all the frames. Finally, using the rotation matrix $\mathbf{R}$ estimated at each frame, we compute the yaw, pitch and roll angles per frame, and represent head rotation throughout the video with a matrix $\boldsymbol{\Phi}$ of size $3 \times T$. The facial movement variation and head rotation of a person throughout the video are represented together with a matrix $\mathbf{Y}$ of size $56 \times T$, obtained as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{E} \\ \boldsymbol{\Phi} \end{bmatrix}. \tag{2}$$

Alternatively, one can consider using the Action Units (AUs) of the Facial Action Coding System instead of the 3DMM-based expression features that we used above. However, our analysis is based on correlation of time series (Section 3.2.2), which requires a representation where AU intensity needs to be provided—binary AU labels would be very limiting. Since automated AU detection systems (*e.g.,* OpenFace [20]) provide AU intensity only for a relatively small number of AUs, we preferred to use the 3DMM-based features instead of the AUs. One could also consider to add the AU features to the features $\mathbf{Y}$ above, but we refrained from doing so, because the number of our correlation features increases exponentially with the number of rows in **Y** (Section 3.2.2). This also explains why we refrained from adding the features from nose and cheek regions, as the potential extra information that would be provided by these regions may not justify the exponential increase in the dimensionality of the feature space. That said, the utility of all such extra information should be explored in future AI pipelines that can be trained with data from larger samples.

3.2.2. Correlation features—

An important aspect of social communication is how different modalities of communicative behavior are integrated and coordinated. For example, the ADOS, the gold standard clinical assessment for autism diagnosis, includes criteria that evaluate how an individual combines speech with gestures and eye contact with facial expression [14]. Similarly, the coordination of behavior within a communicative modality (*e.g.,* movements across different parts of the face) is important; for example, atypical aspects of facial expressions can be characteristic of autism [21, 22]. Thus, to capture coordination across different types of facial and head movements within a person, we apply windowed cross-correlation [23] on the matrix **Y**. That is, considering the ith and jth row of **Y** as two time series, we compute the cross correlation between the two, over time windows of length T_W and a step size of $T_W/2$ (*i.e.,* consecutive time windows have an overlap of 50%). We then compute the average $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$ of the maximal cross-correlation values (*w.r.t.* lag) per window. To distinguish between the cases where, say, a mouth movement was followed with a pose variation from the opposite direction, we allow only forward lag on the second time series in the pair, thus $(\mu_{i,j}, \sigma_{j,i})$ is in general different from $(\mu_{j,I}, \sigma_{j,i})$. In sum, since **Y** has 56 rows, we have 56 × 56 ordered pairs, and with 2 features (*i.e.,* mean and standard deviation) per pair, the total number of features that represent the behavior of a participant is $M = 6272$.

3.2.3. Classification—

We predict the diagnostic group of participants (ASD vs. NT) using a linear SVM classifier by simply using the default C value for SVM (*i.e.,* $C = 1$). We report results based on nested cross-validation, where the only hyper-parameter that is being optimized is the time window T_W, and we optimize over values of $T_W = 1, 2, 4, 6$ seconds. The time window length that was selected in most cross validation folds was $T_W = 2s$.

While more advanced AI models based on deep learning could be used, the sample size is insufficient for reliably training deep learning models from scratch. Moreover, to our knowledge, there is no publicly available pre-trained deep learning model that is directly applicable for our problem, thus taking an existing model and re-training only a part of it (*e.g.,* the classification layer) with our data is also not an approach within reach.

4. Results and Discussion

Table 2 shows the prediction accuracy of the human raters and the AI method. The results for the AI method are obtained via 10-fold cross validation (repeated 100 times with shuffling participant order). The average accuracy of expert clinicians is slightly higher than that of non-experts. Of note, the average accuracy of all human raters (expert and non-expert) is similar to that of the AI approach. The average positive predictive value, negative predictive value, sensitivity and specificity of the AI model are respectively 0.86, 0.79, 0.55, 0.95.

We next investigate whether the errors of the human raters coincide with the errors of the AI algorithm. Table 3 shows the participants whose diagnoses were inaccurately predicted by most human raters (*i.e.,* average prediction accuracy < 50%), along with the correct diagnosis and diagnosis predicted by AI. Results show that four out of these five human mispredictions were correctly predicted by the AI, including the first participant in the list, whose diagnosis was predicted correctly by only 21% of the human raters. In other words, participants that were difficult for most human raters to accurately classify were not particularly difficult for the AI. This suggests that the decision mechanism of AI is different than that of the humans, and the following results further support this point of view.

Fig. 3 plots the average prediction accuracy of human raters against the average accuracy of the AI algorithm per participant. The correlation between these quantities is not strong (ρ = 0.35) and is mostly driven by the participants that are correctly classified by both humans and the AI (*i.e.,* the top right points of the plot). For example, if we remove the subjects that are correctly classified by at least 95% of the human raters, the correlation drops to ρ = 0.19. The lack of points in the lower-left quadrant of the Fig. 3 supports the conclusion that the diagnoses that were difficult to predict for humans were not typically difficult for the AI, and vice versa.

This outcome further supports that the decision mechanism of the AI is different than that of the humans, and is a desirable outcome if AI is to be used as an assistive technology for human clinical decision-making, since it implies that human decisions can be augmented with the help of AI. For example, in a potential application for autism screening from similar short social videos, humans and AI could simultaneously make predictions, and humans could re-evaluate their decision if it is inconsistent with the decision of the AI algorithm. However, arguably, a scenario of this kind is conceivable only if the AI algorithm produces a semantically interpretable output—that is, the algorithm lists the detected behavioral patterns that lead to a diagnostic decision of autism vs. NT. Otherwise, without any explanation of the prediction, it would be difficult for a clinician to determine to what degree the result of the AI algorithm should be taken into account.

In order to shed some light on the decision mechanism of the AI, we analyze the features that were dominant in the SVM classifier—the features that had greater weight. Fig. 4 shows the weights of all the features and Fig. 5a shows the 10 features that had the greatest (absolute) weight across cross-validation folds along with their names. While a complete analysis of the semantic interpretation of each feature is a difficult task, we can still gain some insight into the SVM decisions by inspecting these results. First, note that pose-pose features (*i.e.,* features that summarize correlation between two head rotation angles) have the greatest weight on average (Fig. 4 top), indicating that head movements are important for distinguishing behavioral patterns of autistic vs NT participants. Moreover, correlation features combining the pose and eye emerge as important both in Fig. 4 and in Fig. 5a, supporting previous literature suggesting that blinking and nodding are important non-verbal behaviors in conversations [24], and head and eye movements are indicators of social attention [25]. Second, mouth-related features also emerged as important. For example, six out of 10 correlation features in Fig. 5a are related to mouth, with three of them being pairs of mouth-mouth features.

We next analyze which, if any, of the four feature categories (eyes, brows, mouth, pose) have greater presence among the top *k* features. Fig. 5b plots the proportion of the eye-, brow-, mouth- and pose-related features in the top-10, top-100, top-1000 most important features, as well as their proportion in the entire pool of 6272 features. For example, while the baseline rate of pose features is only ~5.3% (*i.e.,* ~5.3% of the entire set of 6272 features are pose-related), we see that the top 10 features contain a pose-related feature at a ratio of ~13.3% (see caption of Fig. 5 for the computation pose-related features), indicating that the pose features have ~2.5 times more presence in the top-10 features compared to their baseline. Similarly, the baseline rate of mouth-related features is ~25.5%, but ~40% of the top-10 features are related to the mouth, indicating that mouth features also have greater representation in the set of important features compared to their baseline. In sum, our analyses suggest that the AI algorithm places high emphasis on pose- and mouth-related features when classifying between autism and NT groups. Further analysis to uncover why these features are important is beyond the scope of this study, as this would require more granular expression models (*e.g.,* 3D versions of localized bases [26]), because the approach that we designed from an existing model does not allow us to pinpoint the facial movements of interest beyond the level of the partitioned regions in Fig. 2; for example, we cannot distinguish between parts of the mouth, such as upper lip or mouth corner. Still, our analyses allowed a degree of interpretation that corroborates previous findings on the importance of mouth-related movements [2, 4], as well as the central role that head movements have in social orienting, attention and backchannel feedback (*e.g.,* nodding) [27, 28, 24, 25, 29].

5. Conclusions and Future Work

In this paper, we studied the prediction of autism from facial movement behavior during casual conversations. Specifically, we compared the predictive accuracy of expert and non-expert human raters with that of an AI algorithm. Results show that, while both humans and the AI are capable of distinguishing individuals with autism spectrum disorder (ASD) from neurotypical (NT) individuals with high accuracy, their errors do not overlap, suggesting that the decision mechanism of an AI algorithm may be different than that of a human. Thus, AI technologies have the potential to provide complementary information to a clinician and become an assistive tool for decision making. Arguably, the most immediate application based on our results is a new, semi-automatic screening technology for autism, where an individual is advised for further diagnostic evaluation in the event that a (non-expert) human or the AI model predicts that the individual exhibits autism-specific behavior. However, in a real life scenario, the problem of interest would be more difficult as a potential patient may not be NT but may not have ASD either. Thus, future research is needed to identify the performance of humans and AI models in predicting ASD diagnosis from neurodiverse samples.

Our results directly motivate further future research in multiple directions. The most pressing future direction from the perspective of making AI an effective assistive tool is examination of the behaviors that lead to a predicted diagnosis. Having interpretable outputs is necessary for using AI technologies in clinics, as clinicians should understand how the AI algorithm makes a prediction before taking this prediction into account. Furthermore, research on younger participants is needed, given that early diagnosis improves access to effective early interventions and thus can improve developmental outcomes. Another future direction is to investigate the benefits of dyadic analysis, where, unlike our monadic analysis (Section 2), the behavior of confederate is also taken to account. Finally, user research is necessary to test if and to what degree clinician diagnoses can be improved through the use of AI assistive tools.

Acknowledgments

This work is partially funded by the National Institutes of Health (NIH), 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.

References

[1]. Mast MS, Gatica-Perez D, Frauendorfer D, Nguyen L, Choudhury T, Social Sensing for Psy chology: Automated Interpersonal Behavior Assessment, Psychological Science 24 (2015) 154–160. doi:10.1177/0963721414560811.

[2]. Zampella CJ, Bennetto L, Herrington JD, Computer Vision Analysis of Reduced Interpersonal Affect Coordination in Youth With Autism Spectrum Disorder, Autism research : official journal of the International Society for Autism Research 13 (2020) 2133–2142. URL: https://pubmed.ncbi.nlm.nih.gov/32666690/. doi:10.1002/AUR.2334. [PubMed: 32666690]

[3]. Zampella CJ, Sariyanidi E, Hutchinson AG, Bartley GK, Schultz RT, Tunç B, Computational Measurement of Motor Imitation and Imitative Learning Differences in Autism Spectrum Disorder, in: Companion Publication of the International Conference on Multimodal Interaction, Association for Computing Machinery, Inc, 2021, pp. 362–370. URL: 10.1145/3461615.3485426. doi:10.1145/3461615.3485426.

[4]. Parish-Morris J, Sariyanidi E, Zampella C, Bartley GK, Ferguson E, Pallathra AA, Bateman L, Plate S, Cola M, Pandey J, Brodkin ES, Schultz RT, Tunc B, Oral-Motor and Lexical Diversity During Naturalistic Conversations in Adults with Autism Spectrum Disorder, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018, pp. 147–157. URL: http:/aclweb.org/anthology/W18-0616. doi:10.18653/v1/W18–0616.

[5]. McDonald DQ, Zampella CJ, Sariyanidi E, Manakiwala A, DeJardin E, Herrington JD, Schultz RT, Tunç B, Head Movement Patterns during Face-to-Face Conversations Vary with Age, in: Companion Publication of the International Conference on Multimodal Interaction, 2022.

[6]. Levy SE, Mandell DS, Schultz RT, Autism, Lancet 374 (2009) 1627–1638. [PubMed: 19819542]

[7]. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ, DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses., The American journal of psychiatry 170 (2013) 59–70. URL: http://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2012.12070999http://www.ncbi.nlm.nih.gov/pubmed/23111466. doi:10.1176/appi.ajp.2012.12070999. [PubMed: 23111466]

[8]. de Belen RAJ, Bednarz T, Sowmya A, Del Favero D, Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019, Translational psychiatry 10 (2020) 1–20. [PubMed: 32066695]

[9]. Wechsler D, Wechsler Abbreviated Scale of Intelligence - Second Edition (WASI-II), Pearson Clinical, San Antonio, TX, 2011.

[10]. Lord C, Petkova E, Hus V, Gan W, Lu F, Martin DM, Ousley O, Guy L, Bernier R, Gerdts J, Algermissen M, Whitaker A, Sutcliffe JS, Warren Z, Klin A, Saulnier C, Hanson E, Hundley R, Piggot J, Fombonne E, Steiman M, Miles J, Kanne SM, Goin-Kochel RP, Peters SU, Cook EH, Guter S, Tjernagel J, Green-Snyder LA, Bishop S, Esler A, Gotham K, Luyster R, Miller F, Olson J, Richler J, Risi S, A Multisite Study of the Clinical Diagnosis of Different Autism Spectrum Disorders, Archives of General Psychiatry 69 (2012) 306.

URL: http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archgenpsychiatry.2011.148. doi:10.1001/archgenpsychiatry.2011.148. [PubMed: 22065253]

[11]. Lord C, Rutter M, DiLavore PS, Risi S, Gotham K, Bishop SL, Autism diagnostic observation schedule, second edition (ADOS-2), Western Psychological Services, Torrance, CA, 2012.

[12]. APA, Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5, American Psychiatric Association, Washington, D.C, 2013.

[13]. Ratto AB, Turner-Brown L, Rupp BM, Mesibov GB, Penn DL, 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 41 (2011) 1277–86. URL: http://link.springer.com/10.1007/s10803-010-1147-zhttp://www.ncbi.nlm.nih.gov/pubmed/21287253. doi:10.1007/s10803-010-1147-z. [PubMed: 21287253]

[14]. Hus V, Lord C, The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores, Journal of autism and developmental disorders 44 (2014) 1996–2012. [PubMed: 24590409]

[15]. Egger B, Smith WA, Tewari A, Wuhrer S, Zollhoefer M, Beeler T, Bernard F, Bolkart T, Kortylewski A, Romdhani S, et al. , 3d morphable face models—past, present, and future, ACM Trans. Graph 39 (2020) 1–38.

[16]. Cao C, Weng Y, Zhou S, Tong Y, Zhou K, Face-warehouse: A 3d facial expression database for visual computing, IEEE Trans. Vis. Comput. Graph 20 (2013) 413–425.

[17]. Sariyanidi E, Gunes H, Cavallaro A, Automatic analysis of facial affect: A survey of registration, representation, and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2015) 1113–1133. doi:10.1109/TPAMI.2014.2366127. [PubMed: 26357337]

[18]. Sariyanidi E, Zampella CJ, Schultz RT, Tunc B, Inequality-constrained and robust 3d face model fitting, in: Eur. Conf. Comput. Vis, 2020, pp. 433–449.

[19]. Sariyanidi E, Zampella CJ, Schultz RT, Tunc B, Can facial pose and expression be separated with weak perspective camera?, in: IEEE Conf. Comput. Vis. Pattern Recog, 2020, pp. 7173–7182.

[20]. Baltrusaitis T, Zadeh A, Lim YC, Morency L-P, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 59–66.

[21]. Metallinou A, Grossman RB, Narayanan S, 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.

[22]. Guha T, Yang Z, Ramakrishna A, Grossman RB, Hedley D, Lee S, Narayanan SS, 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.

[23]. Boker SM, Rotondo JL, Xu M, King K, Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series, Psychological Methods 7 (2002) 338–355. doi:10.1037/1082-989X.7.3.338. [PubMed: 12243305]

[24]. Gupta A, Strivens FL, Tag B, Kunze K, Ward JA, Blink as you sync: Uncovering eye and nod synchrony in conversation using wearable sensing, in: Proceedings of the 23rd International Symposium on Wearable Computers, 2019, pp. 66–71.

[25]. Foulsham T, Gejdosova M, Caunt L, Reading and misleading: Changes in head and eye movements reveal attentional orienting in a social context, Vision 3 (2019) 43. [PubMed: 31735844]

[26]. Sariyanidi E, Gunes H, Cavallaro A, Learning bases of activity for facial expression recognition, IEEE Transactions on Image Processing 26 (2017) 1965–1978. [PubMed: 28166497]

[27]. Krogsager A, Segato N, Rehm M, Backchannel head nods in Danish first meeting encounters with a humanoid robot: The role of physical embodiment, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8511 LNCS (2014) 651–662. URL: https://link-springer-com.proxy.library.upenn.edu/chapter/10.1007/978-3-319-07230-2_62. doi:10.1007/978-3-319-07230-2_62/COVER.

[28]. Martin KB, Hammal Z, Ren G, Cohn JF, Cas-sell J, Ogihara M, Britton JC, Gutierrez A, Messinger DS, Objective measurement of head movement differences in children

with and without autism spectrum disorder, Molecular Autism 9 (2018) 14. doi:10.1186/s13229-018-0198-4. [PubMed: 29492241]

[29]. Hale J, Ward JA, Buccheri F, Oliver D, Hamilton AFC, Are You on My Wavelength? Interpersonal Coordination in Dyadic Conversations, Journal of Nonverbal Behavior 44 (2020) 63–83. doi:10.1007/S10919-019-00320-3/FIGURES/6. [PubMed: 32189820]

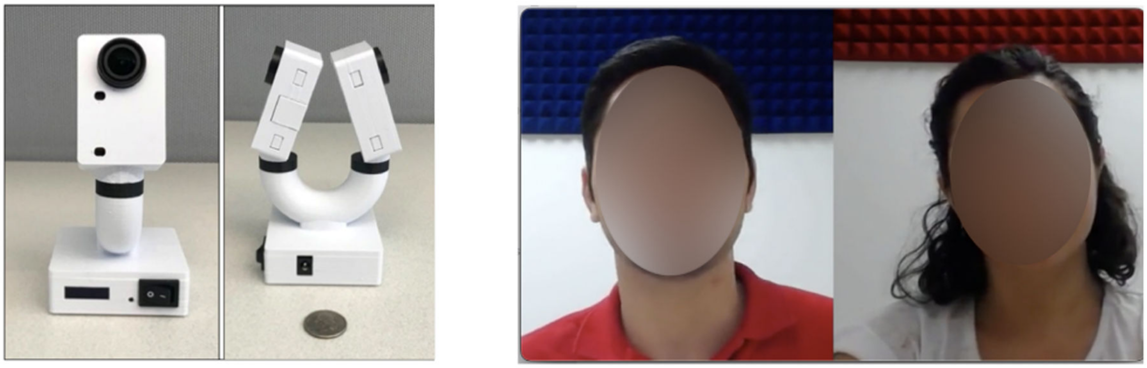

Figure 1:

Left: The device used to record the conversation. The device has two cameras, each pointing to one party of the conversation. Right: Example of videos shown to the human raters. The video contains synchronized videos of the heads/faces of both the participant and the confederate as recorded by the device on the left. The video of the participant’s face only served as input to the AI pipeline.

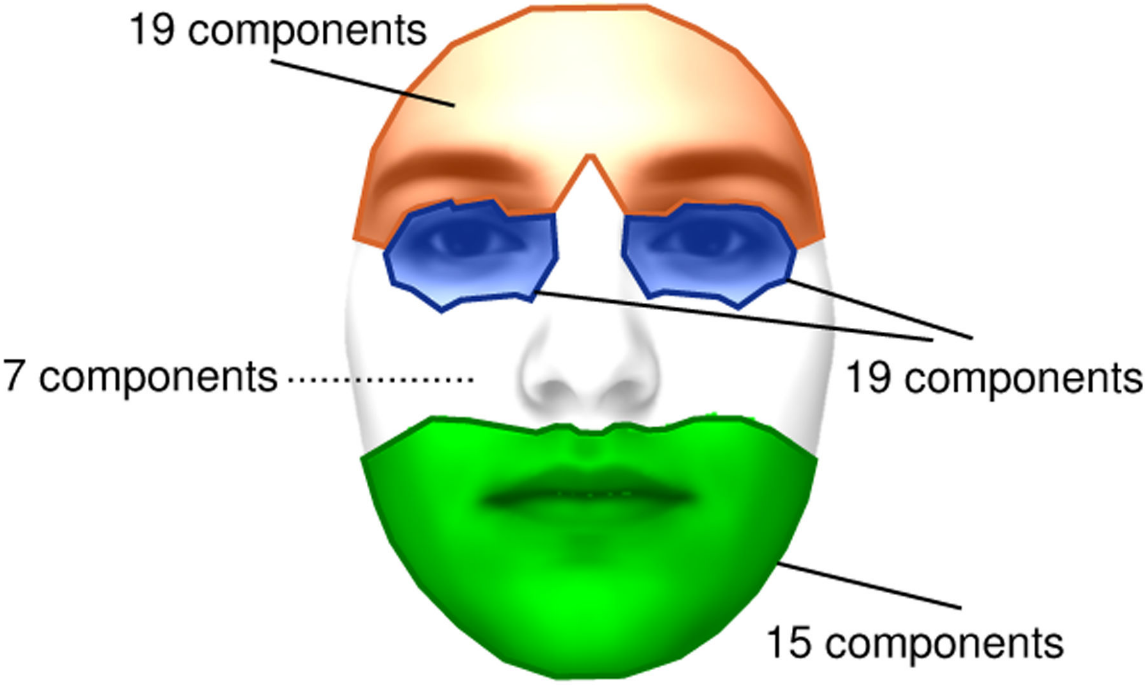

Figure 2:

We divide the facial mesh of *P* points into the four groups illustrated in this figure: (1) brows and forehead; (2) eyes; (3) nose and cheeks; and (4) mouth and chin. Each of the *P* mesh points is assigned to one of these four groups by first computing the distance of the point to all the 51 facial landmarks (iBUG-51 [19]), and then identifying the facial feature (*i.e.,* brow, eye, nose or mouth) corresponding to the closest landmark. The expression basis that we use has a total of 60 components, distributed as shown in the figure.

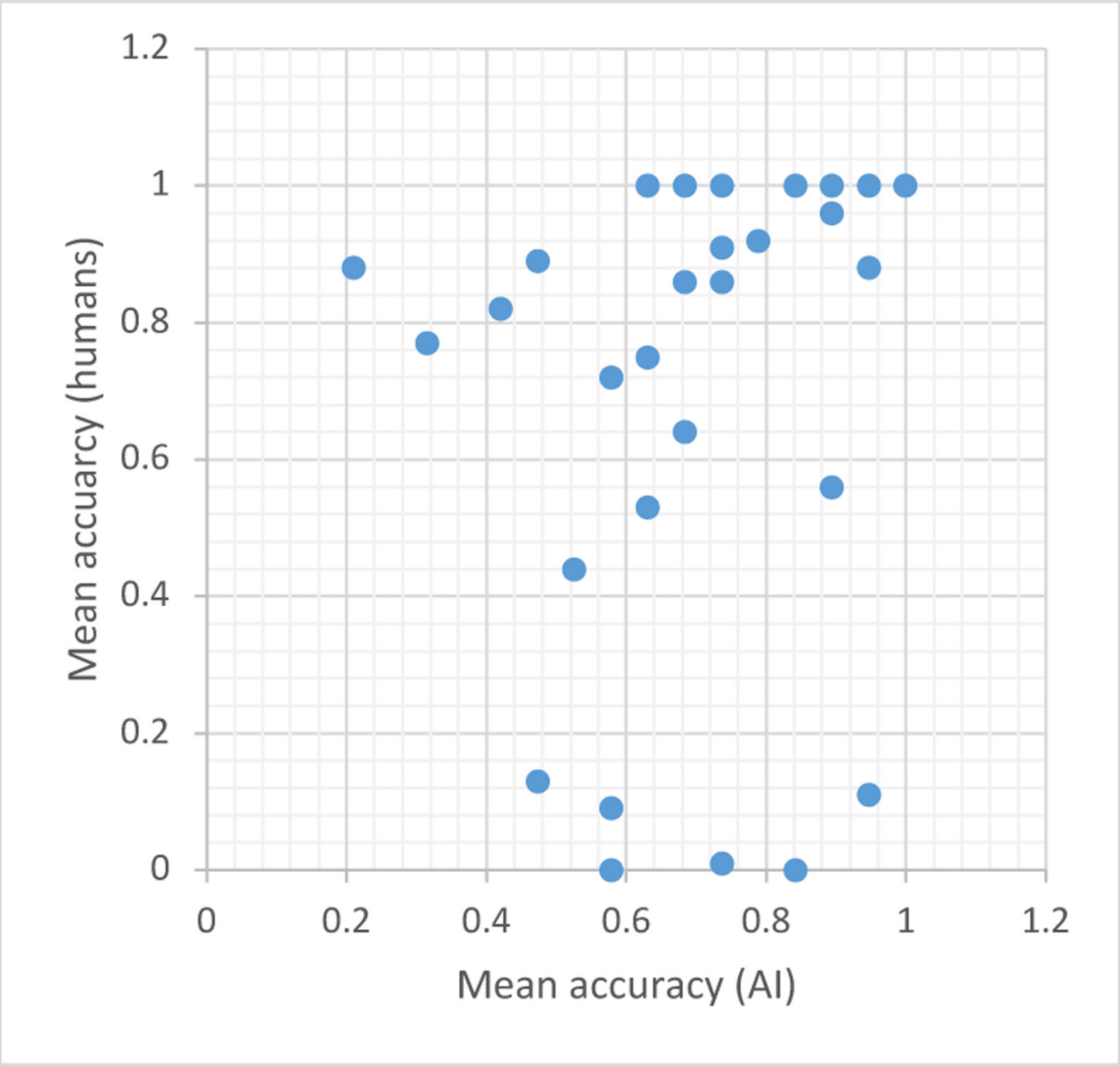

Figure 3:

The average prediction accuracy of human raters against the average prediction accuracy of the AI pipeline, per participant. The average prediction for the AI results in this figure are computed by repeating 5-fold cross-validation 1000 times, and averaging over the predicted 1000 predictions per participant.

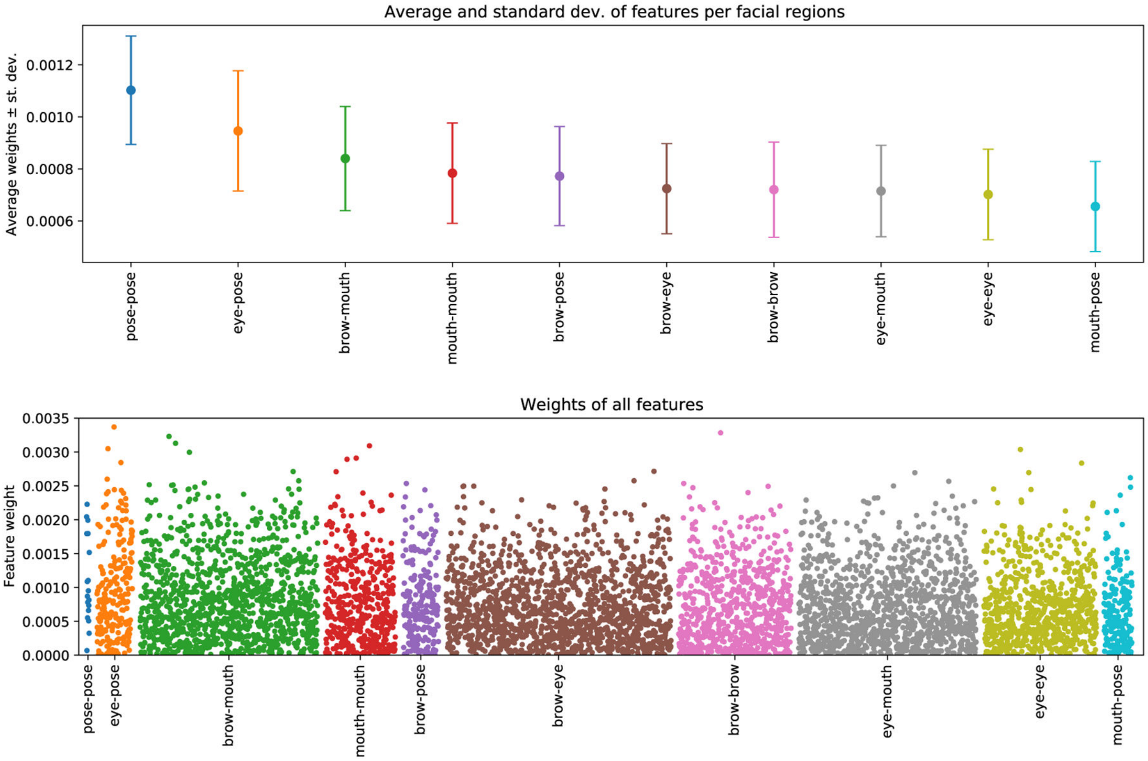

Figure 4:

Top: Average and standard deviation of correlation features per facial region; e.g., statistics for eye-pose show are computed from correlation features that are extracted from these two regions (Section 3.2.2). Bottom: Manhattan plot showing all correlation features.

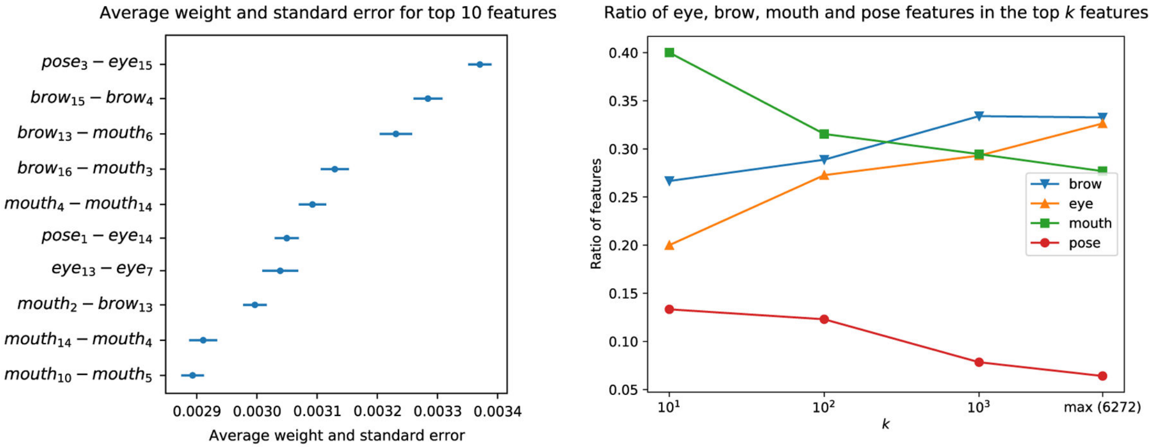

Figure 5:

(a) The labels and weights of the top 10 features along with the standard error (across cross-validation folds). (b) The ratio of each of the four feature categories (brow, eye, mouth, pose) in the top *k* features (*i.e., k* features with the highest average SVM weight) against *k*. The graphs are computed on the basis of a feature category appearing on either side of a correlation feature. For example, if a correlation feature is extracted from the correlation between a mouth and a pose feature, it is considered to be both a mouth and a pose feature. The rightmost value of each graph shows the baseline rate for each feature category –the ratio of the feature category in the entire set of 6272 features– highlighting the importance of the mouth and pose features, since they appear more frequently in the top-10, top-100, top-1000 features compared to their baseline rate.

Table 1

Participant characterization within our sample. Wilcoxon rank sum tests with continuity correction were used for statistical group comparisons, except for sex ratio where the Chi-squared test was used. One NT participant had missing ADOS-2 scores. RRB=Repetitive Behaviors and Restricted Interests subscore of the ADOS-2.

Variable	ASD Mean (SD)	NT Mean (SD)	Statistics	*p*-value
Age (years)	26.9 (7.3)	28.1 (8.4)	W = 234	0.923
Sex (Male, Female)	15m, 2f	23m, 4f	χ^2: 0.08	0.774
Full-Scale IQ	102.1 (19.8)	111.7 (9.5)	W = 157	0.080
Verbal IQ	112.6 (22.1)	112.4(11.2)	W = 215	0.736
ADOS Total	13.1 (3.0)	1.1 (0.9)	W = 442	< 2e-8*
ADOS Social Affect	9.8 (2.3)	1.0 (0.9)	W = 442	< 1e-8*
ADOS RRB	3.3 (1.5)	0.1 (0.3)	W = 441	< 1e-9*

*Statistically significant difference between diagnostic groups, *p*<0.05.

Table 2

Average prediction accuracy of all human raters, non-expert raters, expert raters and AI.

All human raters	Non expert raters	Expert raters	AI
80.3%	78.3%	83.1 *%*	80.5%

Table 3

The five participants whose diagnosis (dx) was mispredicted by most human raters (*i.e.,* average prediction accuracy < 50%), with the corresponding average accuracy by the AI (computed by repeating 5-fold cross-validation 1000 times) and the diagnosis predicted by AI via leave-one-out cross-validation.

dx	Average accuracy (humans)	Average accuracy (AI)	predicted dx (AI, leave-one-out CV)
ASD	21.1%	88.1%	ASD
ASD	31.6%	77.4%	ASD
ASD	42.1%	82.1%	ASD
ASD	47.4%	89.4%	ASD
ASD	47.4%	12.8%	NT