Automatic Identification of High-Risk Autism Spectrum Disorder: A Feasibility Study Using Video and Audio Data Under the Still-Face Paradigm

Chuangao Tang, *Student Member, IEEE*, Wenming Zheng, *Senior Member, IEEE*, Yuan Zong, *Member, IEEE*, Nana Qiu, Cheng Lu, Xilei Zhang, Xiaoyan Ke, and Cuntai Guan, *Fellow, IEEE*

***Abstract*—It is reported that the symptoms of autism spectrum disorder (ASD) could be improved by effective early interventions, which arouses an urgent need for large-scale early identification of ASD. Until now, the screening of ASD has relied on the child psychiatrist to collect medical history and conduct behavioral observations with the help of psychological assessment tools. Such screening measures inevitably have some disadvantages, including strong subjectivity, relying on experts and low-efficiency. With the development of computer science, it is possible to realize a computer-aided screening for ASD and alleviate the disadvantages of manual evaluation. In this study, we propose a behavior-based automated screening method to identify high-risk ASD (HR-ASD) for babies aged 8-24 months. The still-face paradigm (SFP) was used to elicit baby's spontaneous social behavior through a face-to-face interaction, in which a mother was required to maintain a normal interaction to amuse her baby for 2 minutes (a baseline episode) and then suddenly change to the no-reaction and no-expression status with 1 minute (a still-face episode). Here, multiple cues derived from baby's social stress response behavior during the latter episode, including head-movements, facial expressions and vocal characteristics, were statistically analyzed between HR-ASD and typical developmental (TD) groups. An automated identification model of HR-ASD was constructed based on these multi-cue features and the support vector machine (SVM) classifier; moreover, its screening performance was satisfied, for all the accuracy, specificity and sensitivity exceeded 90% on the cases included in this study. The experimental results suggest its feasibility in the early screening of HR-ASD.**

***Index Terms*—High-risk autism spectrum disorder, automated screening, multi-cue features, still-face paradigm, head-movements, facial expressions, vocal characteristics.**

Manuscript received December 26, 2019; revised June 11, 2020 and August 20, 2020; accepted September 22, 2020. Date of publication September 29, 2020; date of current version November 6, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grants 61921004, 61902064, and 81971282, in part by Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX18_0155, in part by the Postdoctoral Science Foundation of China under Grant 2019M661703, and in part by the Fundamental Research Funds for the Central Universities under Grants 2242018K3DN01 and 2242020R20021. *(Corresponding authors: Wenming Zheng; Xiaoyan Ke.)*

Chuangao Tang, Wenming Zheng, Yuan Zong, and Xilei Zhang are with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China (e-mail: tcg2016@seu.edu.cn; wenming_zheng@seu.edu.cn; xhzongyuan@seu.edu.cn; xilei.zhang@seu.edu.cn).

Nana Qiu and Xiaoyan Ke are with the Affiliated Brain Hospital, Nanjing Medical University, Nanjing 210029, China (e-mail: qnn931210@hotmail.com; kexynj@hotmail.com).

Cheng Lu is with the Key Laboratory of Child Development and Learning Science (Ministry of Education), Southeast University, Nanjing 210096, China, and also with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: cheng.lu@seu.edu.cn).

Cuntai Guan is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ctguan@ntu.edu.sg).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2020.3027756

I. Introduction

ASD is a lifelong neurodevelopmental disorder related to impaired social-emotional functioning [1]. The core behavioral symptoms of ASD that appear within two years after birth involve facial expressions, body behaviors and voices, on which the diagnosis of ASD is based [2], [3]. The exact cause of autism is still unclear, and there is no evidence for a cure in the near future [2], but some studies [2], [4], [5] have found that effective early interventions can improve ASD symptoms and outcomes. A delayed diagnosis leads to missing opportunities of early interventions. Therefore, the screening of ASD much earlier than typical diagnosis age at 3-4 years after birth is essential to early interventions. The good news is that some early warning signs before 24 months of age, including less joint attention, lack of social smiles, no response to calling name and communication impairments, etc. [6], [7], have been found in social interactions of babies later diagnosed with ASD. Based on these atypical early symptoms, it is possible to perform an early screening of HR-ASD, which will bring a ray of hope for the babies at risk of ASD.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Currently, the early detection of HR-ASD relies on time-consuming manual measures, including collecting medical history, interviews and behavioral observations. To improve the screening efficiency, an increasing number of researchers focus on developing computer-aided technologies for early identification of ASD [2], [8]. These studies mainly belong to one of two broad categories, including human brain biomarkers and extrinsic behavioral markers. For studies related to the brain, some non-invasive measurements, such as electroencephalography (EEG), magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI), have been employed for finding biomarkers between ASD and healthy comparison groups [9]. Wang *et al.* [3] conducted infant tissue segmentations based on brain MRI scans and performed statistical analyses to identify autistic and normal subjects aged 6 months. Bosl *et al.* [10] proposed using non-linear features, derived from EEG signals, and the SVM calssifier to diagnose HR-ASD cases at 3-36 months of age. On the basis of fMRI signals, Emerson *et al.* [11] defined infants' functional brain connections at 6 months, which was also related to the scores of social behavior, language, motor development and repetitive behavior arising at 24 months of age, and they also used such brain connections as features for identification of HR-ASD. In addition to such automated diagnoses based on costly medical examinations for infants' brains, some researchers proposed to develop behavioral markers-based diagnostic tools [12], where video signals, audio signals and RGB-D (RGB image+depth map) signals captured by low-cost sensors were utilized.

For example, Jaiswal *et al.* [13] designed a paradigm with adult subjects reading and listening to short stories, after which they proposed using computer vision cues derived from RGB-D data as features for detection of ASD and attention-deficit/hyperactivity disorder (ADHD). Liu *et al.* [14] developed a machine learning method for identifying ASD for 4- to 11-year-old children through tracked eye-movement data, which was collected in an experimental scenario where children were asked to distinguish between two races based on facial images. Li *et al.* [15] collected a video-based eye-movement dataset from ASD children (4- to 7-year-old) and TD (6- to 8-year-old) children, and they achieved a diagnostic classification accuracy of 93.7% based on the trajectory of eye movement. Guha *et al.* [16] proposed a computational approach to reveal the facial expressions imitation details at 9-14 years of life for high-functioning autism (HFA) and TD children, where the reduced complexity in dynamic facial behaviors was found to arise primarily from the eye region for those HFA children. Although the existing researches [13]–[16] focusing on an automatic diagnosis of ASD have achieved some progress, yet these studies were based on comparatively older subjects who belonged to groups of children, teenagers or adults. Some aforementioned experimental paradigms and methods are even not applicable to the babies before 24 months of age, because their language skills, behavioral abilities and IQs are still in development. Due to such development gaps, which led to challenges for designing effective behavioral paradigms applicable to babies, the behavior-based automated early screening of HR-ASD was a less-touched problem in the existing researches.

Hashemi *et al.* [17] first designed a mobile application using short movie stimuli to elicit behavioral and social responses from babies, and utilized computer vision algorithms for investigating baby behavioral markers. Sarrett *et al.* [18] applied eye-tracking equipment to study eye fixation in infants later diagnosed with ASD and found that these infants exhibited a mean decline in eye fixation from 2 to 6 months of life. Sheinkopf *et al.* [19] found that HR-ASD infants produced pain-related cries with higher and more variable pitch than those babies in a low-risk group. However, a lack of decision models of binary prediction or severity score is one of common limitations for these markers-related researches, where a final diagnosis can not be provided. Besides, their performance in the scenarios of actual daily social interactions also remains to be seen.

Tronick *et al.* [20] proposed a pioneering paradigm, the still-face paradigm, to assess babies' emotion regulation abilities in actual social interactions. Generally, the still-face paradigm contains 3 episodes, i.e., caregiver-child interaction episode, still-face (SF) episode and reunion episode [20], [21]. The still-face effect has been found robust in most sample variations (infant gender and risk status) and procedural variations (the length of the still-face episodes and the use of intervals between episodes) [21]. A number of studies have employed this paradigm [22] for exploring behavioral markers to further diagnose ASD in adult-baby interaction scenarios. Some initial findings, regarding SF episodes, related to HR-ASD babies before 24 months of age have been achieved, such as more neutral affects [23], fewer frequent gaze shifts [24], longer durations of gazing away from caregiver's face [24], fewer smiles [25], more typical SF effects [26]. Our previous finding [27] showed that babies' social behaviors in the still-face episode were more relevant to the severity of ASD symptoms compared to those in the former mother-baby interaction episode.

To the best of our knowledge, most of the existing SFP-based studies in autism-related fields still undergo the process of manual coding and evaluation. Babies' emotion regulation-relevant cues in the still-face episode, including facial expressions, voices and head-movements, have not been explored for developing automated screening tools to identify HR-ASD.

Overall, the main contributions of this paper are as follows:

1) Multiple vocal and visual features derived from babies' social stress response behaviors were first studied to reveal behavioral differences between HR-ASD and healthy babies aged 8-24 months.
2) A novel behavior-based automated method was proposed for identification of HR-ASD. It has advantages of high-accuracy, low-cost and high-efficiency, and it has potentials for large-scale applications.

II. Data Collection

A. Participants

In this study, 45 infants and toddlers with positive outcomes through the Modified Checklist for Autism in Toddlers

(M-CHAT) screening were preliminarily enrolled to HR-ASD group and 43 typical developmental (TD) infants and toddlers were enrolled to healthy control group. The study was carried out in Nanjing Brain Hospital and was approved by the Medical Ethics Committee of Affiliated Brain Hospital of Nanjing Medical University (2017-KY089-01). All the subjects' guardians agreed that the subjects would participate in this study and signed the informed consent form. For trial registration information, please refer to the Chinese Clinical Trial Registry (ChiCTR-OPC-17011995).

The inclusion conditions for the HR-ASD were as follows: (1) positive screening results based on the M-CHAT; (2) 8 $\leq$ age $<$ 24 months; and (3) the mother was the major caregiver. The exclusion conditions for the HR-ASD consisted of (a) genetic or metabolic disease, such as Rett's syndrome, Fragile X syndrome, etc.; (b) neurodevelopmental disorders, including language developmental disorder, intellectual disability, etc.; (c) traumatic brain injury history; and (d) severe neurological disease history and serious physical illness history.

Participants in the TD group must have met the inclusion conditions of (2) and (3) and all the exclusion criteria as listed for the HR-ASD group.

All participants were assessed with the Gesell developmental schedules [28] at the time of enrollment. To assess the severity of ASD, the babies subjects in the HR-ASD group were assessed with the Communication and Symbolic Behavior Scales Developmental Profile (CSBS-DP) [29], the Childhood Autism Rating Scale (CARS) [30] and the Autism Behavior Checklist (ABC) [31]. Two pediatric psychiatrists provided a final diagnosis based on the Autism Diagnosis Interview-Revised (ADI-R) [32], the Autism Diagnostic Observation Schedule (ADOS) [33] and the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) within one month after their birthdays at 2 years of age.

After re-diagnoses, 5 cases (1 female and 4 males) in the group at risk of ASD were diagnosed with other disorders (language delay) and were categorized to non-ASD group in this study. Limited to the small number of cases with other disorders, a reliable analysis for overall non-ASD group with varying cases could be overgeneralization. Therefore, we narrowed the subsequent analysis to HR-ASD and TD groups.

The demographics of participants in HR-ASD and TD groups are shown in Table I, where the sex of participants was evaluated by the χ^2 test while the age and developmental quotient-based skills were evaluated by the Mann-Whitney U test.

B. Experimental Setup

To capture the data of babies' social behaviors, we employed 4 wireless Ezviz CS-C2C-1B2WFR (1080P) cameras to record videos at a sampling rate of 25 fps. At the same time, the audio data were collected at a sampling rate of 44.1 kHz with a built-in microphone, which is incorporated in a wireless camera device. The experimental scene layout is shown in Fig. S1 that is provided in Supplementary Material.

TABLE I
DEMOGRAPHICS OF PARTICIPANTS (MEAN±SD)

		HR-ASD (n=40)	TD (n=43)	Z/χ^2 value	p-value
	Sex	36(M)/4(F)	32(M)/11(F)	2.47	0.12
	Age(months)	19.65±3.81	16.40±4.70	-3.41	<0.01
Dev. Quotient	Adaptability	78.78±17.07	92.98±7.88	-4.22	<0.01
	Gross Motor	90.95±17.20	92.77±8.46	-0.73	0.47
	Fine Motor	85.83±19.55	93.70±8.29	-2.30	0.02
	Language	58.20±19.84	86.51±8.35	-6.25	<0.01
	Social Skills	78.78±17.35	92.28±7.18	-4.37	<0.01

Notes: SD, standard deviation; M, male; F, female; Dev. Quotient, developmental quotient of the Gesell developmental schedules.

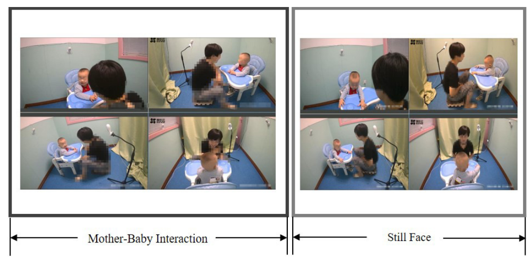

Fig. 1. A snapshot of SFP with two episodes, including an amusing interaction episode (left) and a still-face episode (right).

C. Still-Face Process

In the preparation stage, an experimenter who had assessment experience of babies' behaviors explained the experimental instructions to the mother subject. In the process of face-to-face interaction, the mother subject sat in front of her baby, and the baby subject sat in a baby chair. To avoid unexpected interruptions for the experiment, the experimenter kept quiet and monitored the behavioral experiment from the other side of the same room. At the end of the first episode, the experimenter provided a short voice notice to ask the mother subject to start a new episode.

Following [34], we introduced the SFP by eliminating the reunion episode to make the video and audio data collection procedures more convenient. During the first episode, the mother amused her baby without any touch of body as if at home for 2 minutes. Then, the mother maintained the no-reaction and no-expression status, and placed her gaze above baby's head during the 1-minute still-face episode. A snapshot of our slightly modified SFP procedure is illustrated in Fig. 1.

III. METHODS

In this section, we describe the methods for extracting features from visual and vocal cues. The diagram of our proposed method for the identification of HR-ASD is illustrated in Fig. 2.

A. Head-Movement Feature

To obtain the head-movements features, the OpenPose toolbox [35]–[37] was employed for the estimation of key head points, including the eyes, ears and nose. Among these points, nose point location was found to be more accurate in our preliminary experiment than the other key points.

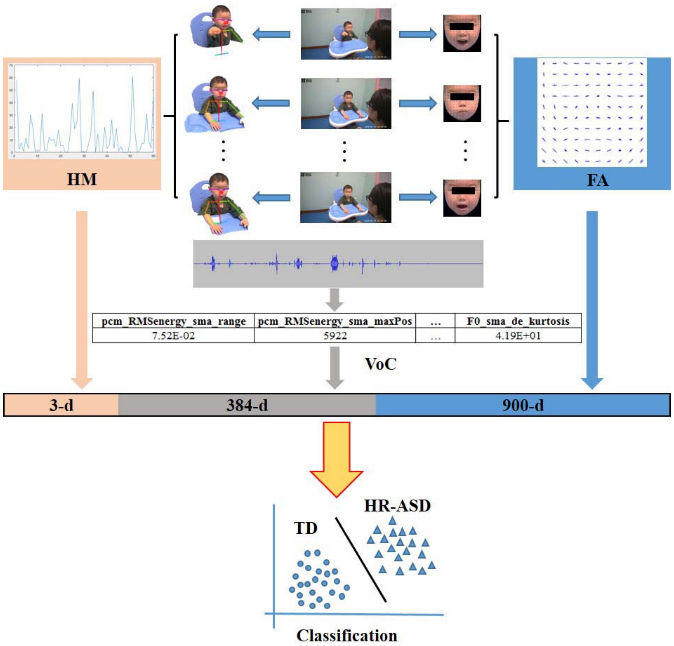

Fig. 2. The proposed automatic method for the identification of HR-ASD. The feature set contains three parts, including 3-dimensional head-movement (HM) features, 384-dimensional vocal characteristics (VoC) features and 900-dimensional HOG-based frame-level average facial appearance (FA) features. Multi-cue-based features were concatenated in serial order to obtain the final fused feature representation for classification.

As a result, the nose point was selected to represent the head center for subsequent head-movement feature analyses.

The babies' atypical head-movements in a social interaction environment could reveal the social impairment of ASD [38]. Here, the babies' head-movements data during the still-face episode were utilized as a distinguishing cue for the classification between the HR-ASD and TD groups. The following statistical indicators for head-movements, including the max-value and mean-value of the head-movement displacement and time delay from the first frame to the frame where the max-value of the head-movement displacement appeared, were computed for analyses.

The point representing the head center for each frame is denoted by $[c_1, c_2, \cdots, c_i, \cdots, c_L]$, where L is equal to the length of the video duration multiplied by its frame rate. Then, the computation of the head-movement-based feature is as follows:

i. Calculate the Mahalanobis distance between c_i and c_1, and then denote the distance vector by $\boldsymbol{d} = [d_1, d_2, \cdots, d_i, \cdots, d_L]$, where $d_1 = 0$;

ii. Calculate the max-value, mean-value of the vector $\boldsymbol{d}$, i.e., max_d_L, $mean\_d_L$, then calculate the time delay δt between the first frame and the frame where the max_d_L appears;

iii. Combine the results into a feature vector $\boldsymbol{v} = (max\_d_L, mean\_d_L, \delta t)$.

B. Facial Appearance Feature

In our experimental scenario, as illustrated in Fig. S1, we set up three cameras to capture baby's facial expressions, i.e., one near-frontal camera and two non-frontal cameras.

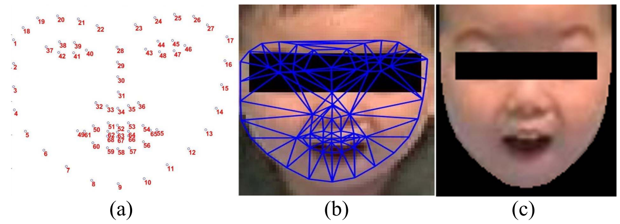

Fig. 3. Visualization for face normalization and masking. From left to right, the images are (a) 68 detected facial landmarks, (b) source: a detected face marked with triangular patches and (c) target: a normalized face with face masking, respectively.

The near-frontal camera aims to capture more facial expression information for favoring the subsequent analysis. As a result, its derived video data were utilized to calculate the facial appearance features.

Some babies showed head-movements during the still-face episode, which resulted in more difficulties for detecting faces, compared with the conditions of the frontal-view facial images. To handle the problem induced by head-movements, we introduced a face detection and alignment toolbox, MTCNN [39], which was designed by deep convolutional neural networks (CNN) and was robust to challenges in unconstrained environments, such as various poses, illuminations and occlusions. The MTCNN toolbox was widely used in the field of face-relevant preprocessing. We re-implemented the face detection framework based on the MTCNN for accurate face locating in sequential frames. The flowchart of re-implementation for MTCNN-based face detection is illustrated in Figure S2, see the Supplementary Material. For small head-movement scenarios, the detected facial region within the predicted bounding box by MTCNN was fed into the OpenFace [40] toolbox for facial image registration as in [41]. First, the toolbox outputted 68 key facial landmarks coordinates for each face, and the face shape can be represented by these points. Then, the current detected face was aligned to the target through a similarity transform, on the basis of the detected facial landmarks and the reference of a frontal facial template [41]. The resolution for a normalized face is 112×112 pixels with a fixed distance of 45 pixels between two pupils. After face normalization, the points surrounding the facial edge were used to mask the face through constructing convex hull. An example for visualization of facial normalization and masking is illustrated in Fig. 3.

Through the face preprocessing as aforementioned, the noise induced by head-movements could be largely reduced for the detected facial images. However, we simply omitted the facial image frame as in [42] for large head-movement scenarios, where the baby's face may not be detected by the face detector.

After face preprocessing, the babies' face detection rates were summarized. Since the face detection rate was not normally distributed, we employed the Mann-Whitney U test to assess significant differences between the two groups. The comparison for face detection rates (mean±sd) corresponding to the HR-ASD and TD groups during the still-face episode is shown in Table II.

Each normalized face was used to calculate the frame-level average facial appearance features. During this process, the facial images were divided into nonoverlapping 12×12

TABLE II
COMPARISON FOR FACE DETECTION RATES OF PARTICIPANTS

HR-ASD	TD	p-value
0.80±0.18	0.84±0.15	0.304

blocks. To alleviate the side effects induced by misregistration error, the blocks on the outermost edge were eliminated and the central 10×10 blocks remained for each facial image.

Some pioneering studies [23], [42] have revealed the facial expressions differences between ASD and non-ASD participants. Here, we further verify this finding by proposing a computational method to detail the differences between the HR-ASD and TD groups.

Human facial expressions are produced by facial muscle deformation according to the well-acknowledged facial action coding system (FACS) [43]. For example, a smile expression is composed of AU6 (Cheeks raised) and AU12 (Lip corners pulled up). Each type of facial muscle deformations corresponds to an unique local facial appearance feature. Motivated by FACS and the development of image descriptors, such as histogram of oriented gradients (HOG), a good representation of appearance and shape information [44], we propose to distinguish the HR-ASD and TD groups through analyzing HOG features that were extracted from local facial regions. It has also shown a more satisfied representation ability than the raw image pixel from the view of better invariance to changes in illumination and shadowing [44]. Concretely, we describe the computation of HOG-based frame-level average facial appearance feature for an image sequence in Algorithm 1, which is presented in Supplementary Material.

C. Vocal Feature

Since the core symptoms of ASD are also involved with voice-related cues [45], we propose to reveal the differences between the HR-ASD and TD groups from the perspective of babies' voices during the still-face episode.

We employed Audacity[1] software for denoising. Both the noise from background and recording device were eliminated as much as possible by the software. Only the baby's voice could be heard after preprocessing.

To quantify the information of voice, low-level descriptors (LLDs) were employed to characterize vocal data from the views of frequency, energy and spectrum. The following sixteen low-level descriptors [45]–[47] were taken into consideration:

- Root Mean Square Energy (RMSE): a characterization that is related to the loudness of a sound signal
- Twelve Mel-Frequency Cepstral Coefficients (MFCC 1-12): a representation of phoneme based on different short-term power spectrums of sound signals [48];
- Zero-Crossing Rate (ZCR) of sound signals: the rate of the signal changing from positive to zero then to negative or vice a verse, which is used for voice activity detection;
- Probability of Voicing (VP): a representation of the probability of detecting the sound signals as voiced;
- Fundamental Frequency (F0): the frequency of vocal chords vibrating in voiced sounds, which is related to prosody.

The LLDs contain two groups of elements, including smoothing of the short-term descriptors (16 elements) and their first-order delta coefficients (16 elements). Twelve statistical functions were computed for these 32 elements to obtain 384-dimensional vocal features (12×32). The employed statistical functions [45], [47] are as follows: arithmetic mean (***amean***), maximum (***max***), minimum (***min***), ***range*** (maximum-minimum), ***maxPos*** (an absolute position corresponding to a maximum value), ***minPos*** (an absolute position corresponding to a minimum value), ***stddev*** (standard deviation), ***slope*** (slope of a linear contour approximation), ***offset*** (offset of a linear contour approximation), ***qerror*** (the quadratic error computed from the actual contour and its linear approximation), ***skewness*** (3th order central moment) and ***kurtosis*** (4th order central moment).

Finally, the vocal feature was extracted through the open-source toolkit openSMILE using the off-the-shelf feature set with aforementioned 384 elements [45], [46].

D. Feature Fusion and Normalization

Motivated by some audio-video-based studies [49], [50], where multiple cues derived from multi-modal signals were fused to attain a better representation, we fused the head-movement, facial appearance and vocal characteristics features to facilitate the improvement of classification performance. At the fusion stage, each unimodal feature vector was concatenated in serial order to attain the final multi-cue representation.

Since each attribute has a different range, it is necessary to conduct column-based feature normalization (samples are represented by row-vectors). The normalization was performed on the training set and then applied to test set. We normalized the value range $[x_{min}, x_{max}]$ to [0,1], and the normalization process can be formulated as follows:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{1}$$

E. Classification and Evaluation

Subsequent to the feature extraction and feature normalization, we employed the support vector machine (SVM) [51] with a linear kernel for classification. SVM seeks a classification hyperplane in a high-dimensional space to separate different types of cases from different categories by maximizing the space between positive and negative groups.

We denote the samples and the corresponding labels as $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_n\}$ and $\{y_1, \cdots, y_i, \cdots, y_n\}$, respectively, where $y_i \in \{-1, +1\}$ and n is the number of samples. The classification hyperplane is as follows:

$$\boldsymbol{w}^T \boldsymbol{x} - b = 0, \tag{2}$$

[1]Audacity® software is copyright ©1999-2019 Audacity Team. Web site: https://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark of Dominic Mazzoni.

TABLE III
CLASSIFICATION RESULTS OF UNIMODAL FEATURES AND MULTI-CUE FEATURES

	SVM					KNN				
	Acc.(%)	Sen.(%)	Spe.(%)	AUC(%)	PPV(%)	Acc.(%)	Sen.(%)	Spe.(%)	AUC(%)	PPV(%)
HM	59.04	90.00	30.23	57.73	54.55	54.22	42.50	65.12	54.07	53.13
FA	89.16	85.00	93.02	89.24	91.89	80.72	65.00	95.35	77.91	92.86
VoC	87.95	90.00	86.05	84.77	85.71	86.75	90.00	83.72	85.17	83.72
HM+VoC	86.75	90.00	83.72	83.14	83.72	87.95	90.00	86.05	85.76	85.71
HM+FA	90.36	85.00	95.35	89.83	94.44	80.72	65.00	95.35	77.91	92.86
VoC+FA	**96.39**	**95.00**	**97.67**	**94.59**	**97.44**	93.98	92.50	95.35	92.50	94.87
HM+FA+VoC	**96.39**	**95.00**	**97.67**	**94.59**	**97.44**	92.77	90.00	95.35	91.40	94.74

where $\boldsymbol{w} = (w_1, w_2, \cdots, w_m)$ is the normal vector of the hyperplane, and b represents the displacement term.

To evaluate the performance of the binary classification, we employed accuracy (Acc.), sensitivity (Sen.), specificity (Spe.), the area under the curve of the receiver operator characteristic (ROC) and the positive predictive value (PPV) as our evaluation indicators. Concretely, sensitivity, namely, the true positive (TP) rate, means the rate of the HR-ASD correctly assigned to the HR-ASD group. Similarly, specificity, namely, the true negative (TN) rate, represents the rate of the subjects in the TD group correctly classified as TD. The value higher than 0.7-0.8 is acceptable for the sensitivity and specificity of a screening tool [52]. The accuracy is computed by $(TP+TN)/N$, where N is the number of all subjects in both groups. ROC is a probability curve, and the AUC provides the distinguishing capability of the classifier between classes, i.e., HR-ASD and TD. Here, the PPV is a probability that subjects with a positive screening test truly have ASD, where the value higher than 0.5 is acceptable [52].

For performance and generalization evaluation, we adopted a subject-independent 10-fold cross-validation protocol to conduct the experiments. In each fold, ~90% subjects were used for training, and the remaining ~10% subjects were tested. We repeated this process 10 times to cover each fold of the data.

To check if the classification accuracy was attained by coincidence, we employed two different classifiers, i.e., SVM (linear kernel, less hyper-parameters compared with other kernels) and KNN (a non-parametric method), for comparison.

IV. RESULTS AND ANALYSIS

A. Classification Results

The classification results corresponding to SVM and KNN classifiers are detailed in Table III. For comparison of screening accuracies between two classifiers, we can find that the SVM classifier (linear kernel) outperformed the KNN (k=5) classifier over all unimodal features. The proposed fusion of three types of features from different modalities shows satisfied performance with all the accuracy, sensitivity, specificity, AUC and PPV exceeding 90% for both classifiers. It also indicates that the fused feature representation is of good discriminability and demonstrates some stabilities for different classifiers.

To evaluate the performance of different kinds of features, we compare each type of feature under the SVM classifier.

TABLE IV
DIAGNOSTIC PREDICTIONS OF MULTI-CUE FEATURES

Actual \ Predicted	TD	HR-ASD
TD	**42**	1
HR-ASD	2	**38**

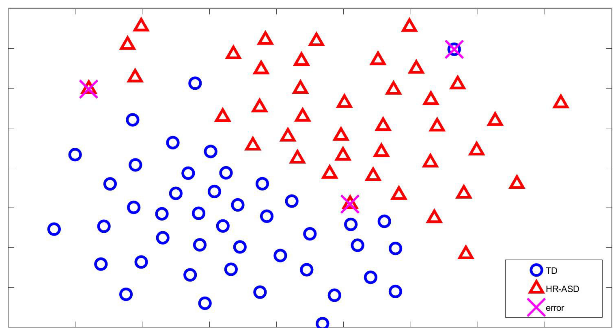

Fig. 4. Visualization for misclassified samples on the 2-D plane by a t-distributed stochastic neighbor embedding.

For unimodal features, the facial appearance (FA) feature achieves the best performance compared with the head-movement (HM) and vocal characteristics (VoC). Fusion of the HM and FA improves the accuracy by ~1.2%, while HM+VoC does not show such an improvement. The fused FA+VoC significantly enhances the accuracy by 7.23% compared with the FA feature which has best classification performance in unimodal field. However, the fusion of FA+VoC+HM does not further improve performance. This may be attributed to the slight contribution of the HM that contains only three statistical elements.

The comparison between diagnostic predictions and actual results of HR-ASD and TD cases is illustrated in Table IV. In total, there were three subjects falsely classified based on the fusion of HM, FA and VoC under the SVM classifier.

In Fig. 4, the misclassified samples are visualized on the 2-D plane through a nonlinear projection. As seen in the Fig. 4, two misclassified HR-ASD samples are close to the samples in the TD group, while the misclassified sample in the TD group seems to be located in the HR-ASD group on the 2-D plane. This may be induced by the comparatively large intraclass covariances for two groups.

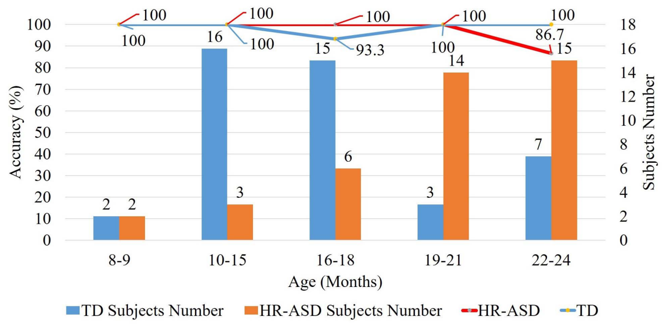

Fig. 5. The classification accuracy and participant number distribution among different month groups.

To assess the statistical significance of the classifier and its classification performance, a permutation test was used. The classification accuracy for each case of 1000 trials (randomly permuting the labels for 1000 times) is presented in Figure S3. Here, the p-value is represented by the proportion of 1000 trials in which the classification performance is the same as or better than the original status under a null hypothesis. From the test results, we find that the classification accuracy corresponding to each case in the permutation test is not higher than the original one before random permutation; thus, the p-value of the permutation test is less than 0.001, which indicates that the alternative hypothesis is true. It implies that the classifier can learn the relationship between the samples' features and corresponding labels. In other words, the multi-cue features can well characterize the discriminant information hidden in the raw video and audio data between the HR-ASD and TD groups.

The recruited subjects in two groups had an overall average age difference of ~3 months (as shown in Table I). The classification accuracy and subject number distribution among different month groups are further analyzed to check if the classification model has a bias due to their ages. As can be seen from Fig. 5, the classification accuracies (blue and red solid lines) are comparatively invariant to age changes. Thus, we can conclude that our identification model does not use age information for classification with a 10-fold cross-validation, where about 10% (8-9/83) independent subjects were used for testing in each fold. Within the same age group (blue and red bar), no category biases can be found because the model does not vote all predictions to TD group or HR-ASD group which contains a comparatively large number of samples. We also find that the proposed method can well predict HR-ASD as early as 8 months of age. The falsely classified samples are in the range between 16 to 18 months and the range between 22 to 24 months, respectively.

B. Statistical Test Analyses for the Extracted Features

We conducted a significant difference test for the extracted features with the Mann-Whitney U test ($\alpha = 0.05$) and employed false discovery rate (FDR) estimation for multiple testing correction.

1) Analyses of Head-Movement Features: The group-level statistical analyses of the head-movement parameters are illustrated in Table V. The results of the U test show that there are no significant differences in max_d_L, $mean\_d_L$ and δt between the HR-ASD and TD groups, respectively; this may be the result of missing social reference-related head-pose information, which is one of the limitations in this study.

TABLE V
STATISTICAL ANALYSIS OF HEAD-MOVEMENT PARAMETERS BETWEEN HR-ASD AND TD (MEAN±SD)

	HR-ASD	TD	p-value
max_d_L	40.90±23.59	37.82±20.22	0.61
$mean\_d_L$	103.56±61.38	99.34±40.90	0.77
δt	33.81±18.20	34.15±16.72	0.80

TABLE VI
GROUP-LEVEL MEAN GRADIENT MAGNITUDES FROM FACIAL REGIONS WITH SIGNIFICANT DIFFERENCES BETWEEN TD AND HR-ASD

Index	TD	HR-ASD	Index	TD	HR-ASD
R006	0.1464	**0.1796**	R007	0.1680	**0.2030**
R008	0.1983	**0.2313**	R019	0.1397	**0.1742**
R020	0.2061	**0.2277**	R021	0.0636	**0.0758**
R022	0.0798	**0.0929**	R026	0.0874	**0.1109**
R027	0.0712	**0.1029**	R028	0.0678	**0.0986**
R031	0.0659	**0.0769**	R032	0.0786	**0.0966**
R036	0.1189	**0.1392**	R037	0.0965	**0.1322**
R038	0.0912	**0.1302**	R039	0.0869	**0.1143**
R042	0.0742	**0.0902**	R043	0.0933	**0.1126**
R047	0.1400	**0.1690**	R048	0.1361	**0.1703**
R049	0.1199	**0.1448**	R052	0.0742	**0.0851**
R053	0.0892	**0.1016**	R079	**0.1737**	0.1456
R080	**0.2301**	0.1838	R081	**0.0948**	0.0762
R086	**0.1613**	0.1343	R087	**0.1636**	0.1334
R088	**0.1900**	0.1514	R089	**0.2350**	0.1885
R091	**0.1275**	0.0934	R092	**0.1486**	0.1080
R093	**0.1821**	0.1326	R094	**0.2105**	0.1566
R095	**0.2309**	0.1714	R096	**0.2414**	0.1786
R097	**0.2564**	0.2006	R098	**0.2714**	0.2315

2) Analyses of Facial Appearance Features: Among the 900 facial appearance features, there are 383 features showing significant differences (FDR-corrected $p < 0.05$). The summation of gradient magnitudes from 9 bins in a histogram for each local facial region was also statistically assessed. As a result, we find that 38 corresponding facial regions (vs. 100 facial regions representing the whole central parts of the face) show significant differences between the two groups (FDR-corrected $p < 0.05$). The group-level mean values for the summation of gradient magnitudes corresponding to these 38 facial regions are shown in Table VI.

Fig. 6 illustrates the visualization for the grayscale frame-level average faces from the HR-ASD and TD groups. From Fig. 6(a)(b), we find that the HR-ASD babies reveal comparatively larger head-poses, which may be induced by a lack of social attention. As for the frame-level average faces, facial expressions from the individuals in the HR-ASD group seem more awkward while most TD babies present expectations or curiosities when their mothers maintain the no-reaction and no-expression status. From Table VI and Fig. 6(c), we find that the group-level mean gradient magnitudes are larger for the right facial regions close to babies'

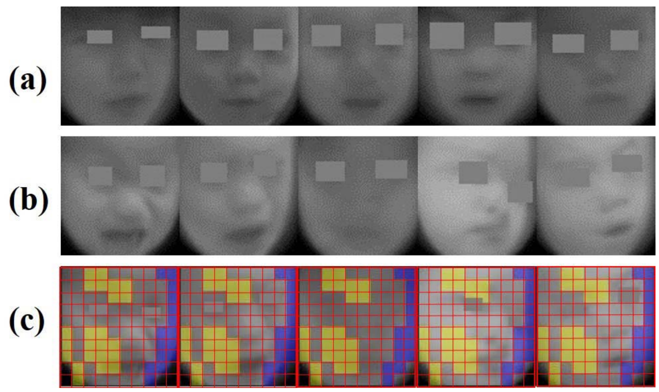

Fig. 6. Visualization for the frame-level average grayscale faces from the TD and HR-ASD groups. (a) TD; (b) HR-ASD; and (c) The highlighted HR-ASD baby facial regions as listed in Table VI. In (c), the regions indexes are with column priority, and yellow color indicates larger gradient magnitudes with blue color for lower gradient magnitudes. This figure should be better viewed in color.

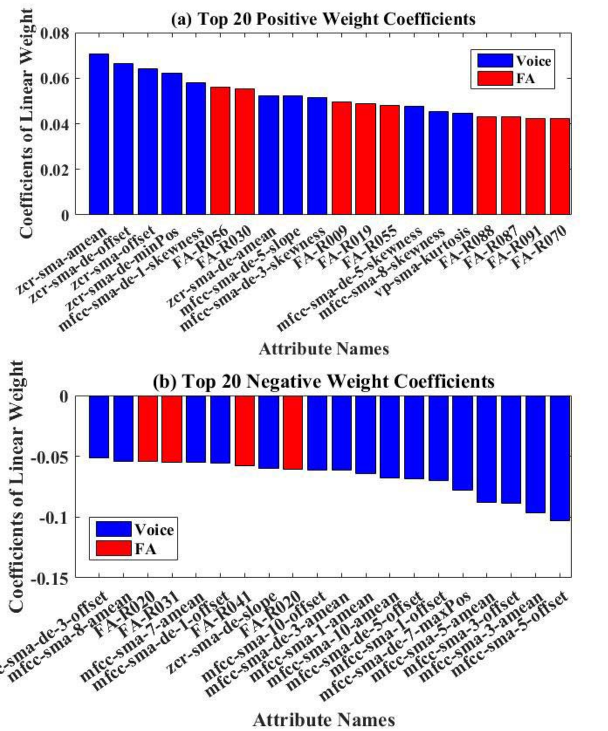

Fig. 7. Visualization for the top 20 positive and negative weight coefficients and the corresponding attribute names. Regarding the appearance feature named FA-R19, the feature is a subtype element of the histogram corresponding to the 19th facial region in Fig. 6. For the vocal features, the features with *'mfcc'* prefix in names are subtype elements corresponding to the MFCC descriptor. The *'sma'* and *'sma_de'* represent smoothing of the short-term descriptors and 1st-order delta coefficients of the smoothed descriptors, respectively. The digital id following 'sma' or 'sma-de' within the *'mfcc'*-related feature names corresponds to the one in 12 Mel-bands. The suffix of *'offset'* in the name is an indication of the corresponding statistical function.

eyes and mouth corner (corresponding to the left part of the image) in the HR-ASD group, this could be induced by HR-ASD babies' awkward facial expressions arising with larger facial muscle deformations. Furthermore, the group-level mean gradient magnitude values are found to be different for partial regions around babies facial edges. One possible explanation is individuals' head-pose differences in the HR-ASD and TD groups.

3) Analyses of Vocal Features: Significant differences between the HR-ASD group versus the TD group can be found from 224 vocal features (FDR-corrected $p < 0.05$). These 224 vocal features are composed of mel-frequency cepstral coefficients (MFCC, 80.8%), root mean square energy (RMSE, 5.8%), zero-crossing rate (ZCR, 8.5%), probability of voicing (VP, 3.1%) and fundamental frequency (F0, 1.8%). F0 (a major cue of prosody)-related features show a significant difference between the HR-ASD and TD groups, which is consistent with the conclusion that those with ASD have problems in prosody [45]. Regarding the rest of the vocal features, including MFCC, ZCR, VP and RMSE, no consistent conclusions have been reached, to the best of our knowledge. In terms of our dataset, we find that most MFCC-based parameters show significant differences between the HR-ASD and TD groups.

C. Visualization of Weights for the Fused Features

The weights denoted in Eq.(2) were computed for visualizing the contribution of each element from the fused features. The top 20 positive and negative weight coefficients and the corresponding features names are illustrated in Fig. 7. As can be seen in this figure, 67.5% of these features belong to the vocal field, and the remaining are related to facial appearance. The results also show that our multi-cue-based method takes advantage of both visual and vocal information.

V. Discussion

In this study, multi-cue features derived from babies' social response behaviors in a frustration environment were statistically analyzed to reveal behavioral differences between HR-ASD and TD groups. The developed multi-cue-based screening method has advantages of high-accuracy, low-cost and noncontact. Different from some pioneering studies [23]–[25], [27], where conventional social behavior indicators under the SF paradigm were manually coded and used for statistical analyses, we proposed a data-driven method that is free of manual coding. This objective measurement, derived from behavioral data, also provides evidence in early screening of HR-ASD. Such a data-driven exploration will inspire researchers from computer vision, pattern recognition and etc. fields to develop more advanced but low-cost behavioral measurement tools in diagnoses of mental disorders.

Limited to the small number of cases with other development disorders in this study, we did not provide a specific analysis for the 5 cases later diagnosed with language delay. Here, a preliminary extension was conducted, and those 5 cases as well as 43 TD cases were merged to non-ASD group for further verification. The diagnostic evaluation for 48 non-ASD cases and 40 HR-ASD cases was conducted through a leave-one-out cross-validation protocol (LOOCV). The SVM (linear kernel) classification model was trained and verified on the proposed multi-cue features, and overall sensitivity, specificity and PPV for total 88 cases were 97.5%(39/40), 89.6%(43/48) and 88.6%(39/44), respectively. Two of the

TABLE VII
COMPARISON BETWEEN RELEVANT SCREENING TOOLS AND OURS

Reference	[53]	[54*]	[55]	**Ours**
Instrument	CHAT	M-CHAT	M-CHAT-R/F	cameras devices
Description	Parent interview and observation	2 stages: Parent questionnaire+Interview	2 stages: Parent questionnaire+Interview	SF paradigm+video recording +automated identification
Age(Months)	18	16-30	16.00-30.95	8-24
Sensitivity	0.35	0.97	0.85	**0.98**
Specificity	0.98	0.95	**0.99**	0.90
PPV	0.59	0.36	0.48	**0.89**

*: the values of sensitivity, specificity and PPV were estimated by Discriminant Function Analysis.

five cases were correctly predicted as non-ASD while the rest 3 cases and 2 TD cases were falsely classified into HR-ASD group. An overall false positive rate of our method is 10.4%(5/48). The comparison between some relevant screening tools and our method is illustrated in Table VII. As can be seen in Table VII, our screening method is appropriate for younger babies than widely used instruments including the checklist for autism in toddlers (CHAT) [53], M-CHAT and the M-CHAT [54], revised with follow-up (M-CHAT-R/F) [55]. The sensitivity of our automated screening method is comparable to M-CHAT, while the PPV seems much better than that of M-CHAT and its modification. Since less negative cases were included in this study compared with [53]–[55], it is still necessary to include a large number of cases for further verification.

Despite the success of the extracted features, there are still opportunities for improving the performance. A lack of robust head-pose measurement for babies' head-movements led this study to using 1st-order indicators for representing head-movement information. The 1st-order statistical analysis for head-movement trajectory is insufficient for understanding atypical social reference. An advanced head-pose estimator may help social reference analysis for babies, which has shown some effectiveness in distinguishing HR-ASD and TD cases under the SF paradigm [6]. Future methods need to incorporate such estimators for further analyses.

Due to a lack of a large number of included cases, this study mainly focused on finding differences between HR-ASD and TD groups. More varying cases with other development disorders were not covered. A large number of cases with matched age need to be included, it could provide opportunities to train a more reliable and robust diagnostic model. The model trained on large-scale samples would be convictive for medical community, and other researchers can employ the off-the-shelf diagnostic model for more explorations. Moreover, it is significant to include more younger babies earlier than 8 months of age, and it will reveal the earliest age when the automated screening method could provide an acceptable diagnostic result.

In order to be applicable to unconstrained environments including homes and child health care centers, future work should refine the experimental layout, e.g., an example video for guiding participants how to perform under the paradigm should be incorporated. The proposed method should be extended to an end-to-end system which could be installed on some smart devices for large-scale applications.

VI. CONCLUSION

This paper presents a multi-cue-based automated screening method for early identification of infants and toddlers at high risk for ASD before 24 months of life. Under the simple but effective still-face paradigm, multiple features derived from babies' visual and vocal behavior were analyzed to reveal differences between HR-ASD and TD. The proposed multi-cue features showed better diagnostic performance than the unimodal features, which verifies the effectiveness of our proposed method. Such an automated identification tool could meet the need of large-scale screening for ASD.

REFERENCES

[1] W. Zheng *et al.*, "Multi-feature based network revealing the structural abnormalities in autism spectrum disorder," *IEEE Trans. Affect. Comput.*, early access, Jan. 1, 2019, doi: 10.1109/TAFFC.2018.2890597.

[2] J. Liu, K. He, Z. Wang, and H. Liu, "A computer vision system to assist the early screening of autism spectrum disorder," in *Proc. Int. Conf. Cogn. Syst. Signal Process.* Singapore: Springer, 2018, pp. 27–38.

[3] L. Wang *et al.*, "Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 411–419.

[4] X. Liu, Q. Wu, W. Zhao, and X. Luo, "Technology-facilitated diagnosis and treatment of individuals with autism spectrum disorder: An engineering perspective," *Appl. Sci.*, vol. 7, no. 10, p. 1051, Oct. 2017.

[5] Z. Warren, M. L. McPheeters, N. Sathe, J. H. Foss-Feig, A. Glasser, and J. Veenstra-VanderWeele, "A systematic review of early intensive intervention for autism spectrum disorders," *Pediatrics*, vol. 127, no. 5, pp. 1303–1311, May 2011.

[6] N. Qiu *et al.*, "Application of the still-face paradigm in early screening for high-risk autism spectrum disorder in infants and toddlers," *Frontiers Pediatrics*, vol. 8, p. 290, Jun. 2020, doi: 10.3389/fped.2020.00290.

[7] J. Barbaro and C. Dissanayake, "Early markers of autism spectrum disorders in infants and toddlers prospectively identified in the social attention and communication study," *Autism*, vol. 17, no. 1, pp. 64–86, Jan. 2013.

[8] M. Fakhoury, "Autistic spectrum disorders: A review of clinical features, theories and diagnosis," *Int. J. Develop. Neurosci.*, vol. 43, no. 1, pp. 70–77, Jun. 2015.

[9] J. Kang, H. Chen, X. Li, and X. Li, "EEG entropy analysis in autistic children," *J. Clin. Neurosci.*, vol. 62, pp. 199–206, Apr. 2019.

[10] W. J. Bosl, H. Tager-Flusberg, and C. A. Nelson, "EEG analytics for early detection of autism spectrum disorder: A data-driven approach," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 6828.

[11] R. W. Emerson *et al.*, "Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age," *Sci. Transl. Med.*, vol. 9, no. 393, Jun. 2017, Art. no. eaag2882.

[12] J. Thevenot, M. B. Lopez, and A. Hadid, "A survey on computer vision for assistive medical diagnosis from faces," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1497–1511, Sep. 2018.

[13] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley, "Automatic detection of ADHD and ASD from expressive behaviour in RGBD data," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 762–769.

[14] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Res.*, vol. 9, no. 8, pp. 888–898, Aug. 2016.

[15] J. Li, Y. Zhong, and G. Ouyang, "Identification of ASD children based on video data," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 367–372.

[16] T. Guha, Z. Yang, R. B. Grossman, and S. S. Narayanan, "A computational study of expressive facial dynamics in children with autism," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 14–20, Jan. 2018.

[17] J. Hashemi *et al.*, "Computer vision analysis for quantification of autism risk behaviors," *IEEE Trans. Affect. Comput.*, early access, Sep. 3, 2018, doi: 10.1109/TAFFC.2018.2868196.

[18] J. C. Sarrett and K. S. Rommelfanger, "Commentary: Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, p. 427, Dec. 2015.

[19] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder," *Autism Res.*, vol. 5, no. 5, pp. 331–339, 2012.

[20] E. Tronick, H. Als, L. Adamson, S. Wise, and T. B. Brazelton, "The Infant's response to entrapment between contradictory messages in face-to-face interaction," *J. Amer. Acad. Child Psychiatry*, vol. 17, no. 1, pp. 1–13, Dec. 1978.

[21] J. Mesman, M. H. van IJzendoorn, and M. J. Bakermans-Kranenburg, "The many faces of the still-face paradigm: A review and meta-analysis," *Develop. Rev.*, vol. 29, no. 2, pp. 120–162, Jun. 2009.

[22] L. Giusti, L. Provenzi, and R. Montirosso, "The face-to-face still-face (FFSF) paradigm in clinical settings: Socio-emotional regulation assessment and parental support with infants with neurodevelopmental disabilities," *Frontiers Psychol.*, vol. 9, p. 789, May 2018, doi: 10.3389/fpsyg.2018.00789.

[23] N. Yirmiya, I. Gamliel, T. Pilowsky, R. Feldman, S. Baron-Cohen, and M. Sigman, "The development of siblings of children with autism at 4 and 14 months: Social engagement, communication, and cognition," *J. Child Psychol. Psychiatry*, vol. 47, no. 5, pp. 511–523, May 2006.

[24] L. V. Ibanez, D. S. Messinger, L. Newell, B. Lambert, and M. Sheskin, "Visual disengagement in the infant siblings of children with an autism spectrum disorder (ASD)," *Autism*, vol. 12, no. 5, pp. 473–485, Sep. 2008.

[25] T. D. Cassel, D. S. Messinger, L. V. Ibanez, J. D. Haltigan, S. I. Acosta, and A. C. Buchman, "Early social and emotional communication in the infant siblings of children with autism spectrum disorders: An examination of the broad phenotype," *J. Autism Develop. Disorders*, vol. 37, no. 1, pp. 122–132, Jan. 2007.

[26] S. Ostfeld-Etzion, O. Golan, Y. Hirschler-Guttenberg, O. Zagoory-Sharon, and R. Feldman, "Neuroendocrine and behavioral response to social rupture and repair in preschoolers with autism spectrum disorders interacting with mother and father," *Mol. Autism*, vol. 6, no. 1, p. 11, 2015.

[27] N. Qiu *et al.*, "A study of the makers of early social behaviors of toddlers with high-risk autism spectrum disorder," *Chin. J. Psychiatry*, vol. 52, no. 1, pp. 50–56, 2019.

[28] W. K. Frankenburg and J. B. Dodds, "The denver developmental screening test," *J. Pediatrics*, vol. 71, no. 2, pp. 181–191, Aug. 1967.

[29] A. M. Wetherby and B. M. Prizant, *Communication and Symbolic Behavior Scales: Developmental Profile*. Baltimore, MD, USA: Paul H Brookes, 2002.

[30] E. Schopler, R. J. Reichler, R. F. DeVellis, and K. Daly, "Toward objective classification of childhood autism: Childhood autism rating scale (CARS)," *J. Autism Develop. Disorders*, vol. 10, no. 1, pp. 91–103, Mar. 1980.

[31] D. A. Krug, J. Arick, and P. Almond, "Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior," *J. Child Psychol. Psychiatry*, vol. 21, no. 3, pp. 221–229, Jul. 1980.

[32] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *J. Autism Develop. Disorders*, vol. 24, no. 5, pp. 659–685, Oct. 1994.

[33] C. Lord *et al.*, "The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Develop. Disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[34] L. Provenzi, R. Borgatti, G. Menozzi, and R. Montirosso, "A dynamic system analysis of dyadic flexibility and stability across the Face-to-Face still-face procedure: Application of the state space grid," *Infant Behav. Develop.*, vol. 38, pp. 1–10, Feb. 2015.

[35] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008

[36] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[37] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[38] K. B. Martin *et al.*, "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Mol. Autism*, vol. 9, no. 1, p. 14, 2018.

[39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[40] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 59–66.

[41] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–6.

[42] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, "A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 353–361, Feb. 2018.

[43] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K.: Oxford Univ. Press, 1997.

[44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[45] N. Russo, C. Larson, and N. Kraus, "Audio–vocal system regulation in children with autism spectrum disorders," *Exp. Brain Res.*, vol. 188, no. 1, pp. 111–124, 2008.

[46] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[47] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 835–838.

[48] H. Skowronski, "Increased MFCC filter bandwidth for noise-robust phoneme recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2002, p. 801.

[49] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, Oct. 2018.

[50] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 3–10.

[51] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[52] R. E. Nickel and L. Huang-Storms, "Early identification of young children with autism spectrum disorder," *Indian J. Pediatrics*, vol. 84, no. 1, pp. 53–60, Jan. 2017.

[53] G. Baird *et al.*, "A screening instrument for autism at 18 months of age: A 6-Year follow-up study," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 39, no. 6, pp. 694–702, Jun. 2000.

[54] M. K. Khowaja, A. P. Hazzard, and D. L. Robins, "Sociodemographic barriers to early detection of autism: Screening and evaluation using the M-CHAT, M-CHAT-R, and follow-up," *J. Autism Develop. Disorders*, vol. 45, no. 6, pp. 1797–1808, Jun. 2015.

[55] D. L. Robins, K. Casagrande, M. Barton, C.-M.-A. Chen, T. Dumont-Mathieu, and D. Fein, "Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F)," *Pediatrics*, vol. 133, no. 1, pp. 37–45, Jan. 2014.