2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Responsive Social Smile: A Machine Learning based Multimodal Behavior Assessment Framework towards Early Stage Autism Screening

Yueran Pan
Duke Kunshan University
Kunshan, China
yueran.pan@dukekunshan.edu.cn

Kunjing Cai
Sun Yat-sen Universit
Guangzhou, China
caikj3@mail2.sysu.edu.cn

Ming Cheng
Duke Kunshan University
Kunshan, China
ming.cheng@dukekunshan.edu.cn

Xiaobing Zou
The Third Affiliated Hospital, Sun Yat-sen University
Guangzhou, China

Ming Li
Duke Kunshan University
Kunshan, China
ming.li369@dukekunshan.edu.cn

***Abstract*—Autism spectrum disorder (ASD) is a neuro-developmental disorder, which causes deficits in social lives. Early screening of ASD for young children is important to reduce the impact of ASD on people's lives. Traditional screening methods mainly rely on protocol-based interviews and subjective evaluations from clinicians and domain experts, which requires advanced expertise and intensive labor. To standardize the process of ASD screening, we design a "Responsive Social Smile" protocol and the associated experimental setup. Moreover, we propose a machine learning based assessment framework for early ASD screening. By integrating speech recognition and computer vision technologies, the proposed framework can quantitatively analyze children's behaviors under well-designed protocols. We collect 196 stimulus samples from 41 children with an average age of 23.34 months, and the proposed method obtains 85.20% accuracy for predicting stimulus scores and 80.49% accuracy for the final ASD prediction. This result indicates that our model approaches the average level of domain experts in this "Responsive Social Smile" protocol.**

I. Introduction

Autism spectrum disorder (ASD) is a mental disorder that may significantly impact people's lives. Difficulties in social communication represent one category of the autistic syndromes [1]. Among the indicators that can reflect our sensing and understanding of the social environment, the smile is one of the most important factors. Especially for children between 1 to 3 years old, their responses to simple utterances, including a responsive social smile, can reflect the level of ASD in some widely used screening tests (e.g., ASD-G [2], ADOS-2 [3]).

Although social smiling is essential in ASD screening, most existing methods still rely on clinicians' subjective evaluation. According to screening sheets, clinicians can distinguish ASD and evaluate multi-dimensional abilities. Unlike other screening methods involving clear indicators and benchmarks (e.g., response to name, joint attention), the responsive social smile is evaluated by emotion description that usually remains an ambiguous classification result. It is especially harder for clinicians to assess children's emotions on small faces and record them when distracted from carrying out experiments. Remembering every score for different social stimuli during the screening or perform the behavior scoring after the test is very subjective and labor-intensive, which requires experience and expertise. As top clinicians master the scoring skills with high accuracy, we are motivated to imitate experts and develop an intelligent, objective, and efficient system that can recognize social stimuli, detect and quantify the corresponding emotion in ASD screenings. In this case, clinicians would be equipped with leading technology and rich experience and universally serve autism children.

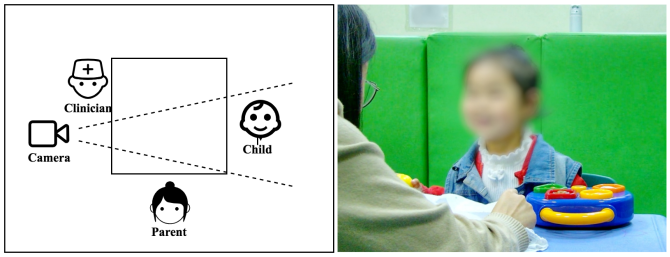

Fig. 1. The layout of the experimental environment and video recording example.

First, we design a user-friendly protocol to standardize the experiments of responsive social smiling. Second, we employ a set of audio-visual pattern recognition modules, including speech recognition, face recognition, and facial expression recognition, to measure children's responses to social stimuli. Third, based on the multimodal behavioral features extracted from the original audio-visual data, we utilize a machine learning based classifier to predict the ASD label, which is usually evaluated by clinicians.

The first two authors contributed equally.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

TABLE I
COMPARISONS OF TYPICAL METHODS

Authors	Method	Algorithm	Accuracy	Sensitivity	Specificity	Data Scale (ASD/Non-ASD)	Age (Years)
Liu et al. [4]	Eye movement	K-means + SVM	88.51%	93.10%	86.21%	29/58	4-11
Li et al. [5]	Hand imitation tasks	Linear SVM	86.70%	85.70%	87.50%	16/14	2-4
Nakai et al. [6]	Abnormal prosody	SVM	76.00%	81.00%	73.00%	31/51	3-10
Heinsfeld et al. [7]	Neuroimaging	Neural Networks	70.00%	74.00%	63.00%	505/535	7-64
Ours	Responsive social smile	CNN + Decision Tree	80.49%	85.00%	77.27%	20/21	1-3

In the database collected by our collaborative hospital, our automatic assessment framework achieves the accuracy of 85.20% for stimulus scoring and 80.49% accuracy for ASD classification. The assessment framework allows clinicians to concentrate on experiments, delivers standardized evaluation with readable scores, and provides an additional classification on ASD. It is worth noting that it generates complementary and comparable performance against other widely used tasks (e.g., response to name, joint attention). Although using the responsive social smile score alone cannot give a very comprehensive ASD classification, we believe that fusing scores and features from multiple automatic assessment tasks can further enhance the overall ASD screening performance.

II. Related Work

Traditional methods for screening ASD often require a series of structured tests involving social interactions between the clinician and the child under the assessment. In the field of psychology, two golden standards are Autism Diagnostic Observation Schedule-Generic (ADOS-G) [8] and the revised version ADOS-2 [2]. Clinicians also want to investigate children's performance in their daily lives by questionnaires [9], [10], [11]. The parents of children under assessment need to finish a questionnaire with multi-choice questions, e.g., motion and voice imitation, stereotyped and repeated action, environmental perception, social communication, language development. Both ADOS screening and questionnaire are subjective and have high demands on experienced clinicians.

In recent years, many researchers begin to explore the feasibility of applying machine learning based methods to perform assistive screening. The goal is to design intelligent algorithms for observing and analyzing children's social behaviors quantitatively. Kosmicki et al. [12] utilize SVM to investigate the potential importance of different behavioral scores in ADOS-G [8] screening. Thus, adopting a subset of all the scores can reduce the time cost in screening processes. It shows that the most relevant scores are usually generated from children's voice, visual attention, hand movement, emotion, and other behaviors, reflecting social abilities.

Based on the above results, some researchers turn to recognize ASD scores and detect autism from behavior data by machine learning methods. There are already some powerful tools that are easy to use. In the speech processing domain, Kaldi [13] is one of the most popular speech recognition toolkits, which can work well in many applications. In the field of computer vision, OpenCV is popular in digital image processing. Dlib [14] is another useful tool, it contains a set of CNN models for face detection (e.g., MMOD [15]).

For observed behavior data, Jiang et al. [16] present a neural network model to learn human visual attention, then perform an ASD assessment. Liu et al. [3], [4] propose methods to identify ASD children by gaze patterns. In screening experiments, 4-11 years old children are asked to recognize faces in pictures, and their eye fixation trajectories are recorded. The recorded eye movement data are partitioned into regions by the k-means algorithm and then represented by histogram-based feature extraction [17]. Finally, an ASD classification accuracy of 89.63% is reported by using SVM as the classifier. Li et al. [5] analyze hand imitation tasks recorded by the motion tracker. The movement data is from 30 adults and described by 20 kinematic parameters. In the paper, they try four different machine learning classifiers (SVM, Random Forest, Naive Bayes, and Decision Tree) to predict ASD, with the best accuracy of 86.70%. Hashemi et al. [18] design a mobile application, which presents fixed movie stimuli to children and records their reaction to screen ASD. This mobile application reflects various indicators, but it only implements two screening protocols for toddlers: response to name and emotion recognition. It shows that the ASD group's response latency is significantly longer compared to the non-ASD group. It also reports that ASD children exhibit fewer positive emotions than non-ASD children when watching the same movies.

Based on voice analysis, Nakai et al. [6] utilize abnormal prosody to detect the ASD. In their experiments, 3-10 years old children are asked to recognize picture cards under instructions verbally. Children's voice is to be judged by speech therapists and a machine learning method. The authors conclude that the SVM classifier performs better than humans. Li et al. [19] present a method to evaluate children's atypical prosody and stereotyped idiosyncratic phrases by speech processing methods.

Rather than screening autistic children just from a single aspect, the latest research tries to build a complex assessment framework that can detect autism from multimodal data. Liu et al. [20] propose a regularized procedure-related assessment framework for response to name and questionnaire research.

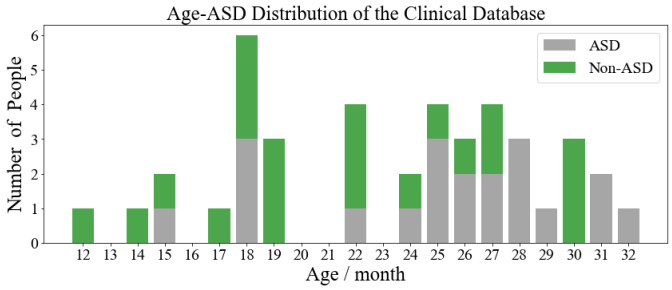

Fig. 2. Age-ASD distribution in the clinicial database.

In the framework, the algorithm needs the start time of a name call whether the child turns his head, looks at the voice source, and records the child's performance. The framework integrates a series of technologies involving speech recognition, face detection, and face verification. The interaction from name call shows children's understanding of a single word rather than conveyed emotion. It would be better to choose other protocols to analyze children's understanding and expression in social communication.

Moreover, many scientists applied advanced physical and biological technologies, e.g., nuclear magnetic resonance and genetic testing for ASD screening. Heinsfeld et al. [7] explore the ABIDE [21] database that contains brain neuroimaging data from 7-64 years old people, and they achieve 70.00% accuracy using DNN. However, applied equipment is too expensive to be prevalent. It is also difficult for young children to stay still and be scanned by neuroimaging devices.

Some widely used methodologies are shown in Table I. It shows that cost and age requirements prevent some accurate ASD screening methods among children. How to automatically evaluate young children's social understanding remains a challenging task. To solve the problem, we present an automatic framework to detect the behavior score in the psychological experiment of the responsive social smile [2]. Our contributions can be summarized as follows:

- We design standardized test environments with low-cost devices and a structured protocol.
- We deploy the hardware and protocol in our collaborative hospital and collect a multimodal database containing 41 young children. Professional doctors diagnose all these 41 children and provide detailed behavior scores as well as ASD labels.
- We build a large video-based facial expression database dedicated to children under six years old. This database has 15,000 video clips from different children under six years old, and it is used to fine-tune our expression recognition module. It is specially set up for young children who are diagnosed with autism, while others are not.
- The proposed framework could provide an overall ASD classification and the corresponding behavior scores of various stimuli, which helps clinicians' diagnose and intervention.

TABLE II
STIMULI AND KEY WORDS IN A PROTOCOL.

	Stimulus	Key Words	Voice Source
1	Greeting smile	"Hello!" + Children's names	Clinician
2	Praise words	"You are so cute/cool!"	Clinician
3	Hide and seek	"Let's play hide and seek'.'	Clinician
4	Hints of tickling	"I am going to tickling you!"	Clinician
5	Tickling	"Real tickling now!"	Clinician
6	Greeting smile	"Hello!" + Children's names	Parent

- Lastly, a machine learning based classifier is employed to give the final classification result for ASD screening.

III. PROTOCOL AND DATABASE

In the experiment, a child, a clinician, and a parent sit as the layout in Fig. 1. A camera can directly record the child's front face. The participating child's behaviors are evaluated as 0 or 1 for each stimulus by a clinician. To be more specific, 0 represents a clear response and smile, which means that the child could hear the stimulus, understand the social meaning, and express emotional responses. While score 1 represents no responses for the given stimulus. Whether the child looks back to the stimulus source or becomes happy can reflect his understanding and social response level. Because sometimes a child does not participate in every step of the social smile experiments, some data do not contain all kinds of listed stimuli.

A. Procedure of the responsive social smile protocol

Taking different humor mechanisms into consideration and following the ADOS design, we introduce five types of social stimuli into our protocol: greeting smile, praise words, hide and seek, hints of tickling, and tickling. There are three participants in an experiment: a child, a parent, and a professional clinician. The clinician would give several or all kinds of stimuli. Sometimes parents would try an additional stimulus of praise words. Therefore, there are six stimuli together, as shown in Table II.

During each step, a clinician or a parent would say key words to indicate the start time of each stimulus. When a child does not respond to words or behaviors, the clinician or the parent would repeat the stimulus three times at most.

B. Clinical Database

By collaborating with the Third Affiliated Hospital of Sun Yat-sen University, we have recorded multimodal data of this protocol with 41 children (shown in Figure 2). Professional clinicians diagnose that 20 are labeled as ASD, while the rest 21 are labeled as non-ASD. According to the DSM-IV [22] standards, those young children are from 12 to 32 months old, and the average age is 23.34 months. There are 33 boys and 8 girls, of which the male to female ratio is skew distributed but close to 4:1, matching with the common gender ratio of ASD [23].

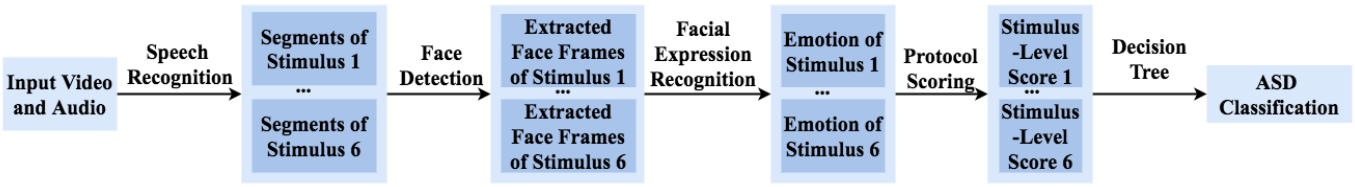

Fig. 3. The proposed assessment framework for analyzing the "Responsive Social Smile" protocol.

IV. Multimodal Assessment Framework

Figure 3 illustrates the whole framework for "Responsive Social Smile" consisting of five stages: temporal stimulus localization, face detection, facial expression recognition, stimulus scoring, and ASD classification. This framework is designed to standardize the experimental procedures. To be more specific, speech and image processing methods are applied to quantify each step of the developed protocol and give stimulus scores. Based on the feature vector composed of stimulus scores, we utilize a decision tree classifier to predict whether a participant has ASD or not.

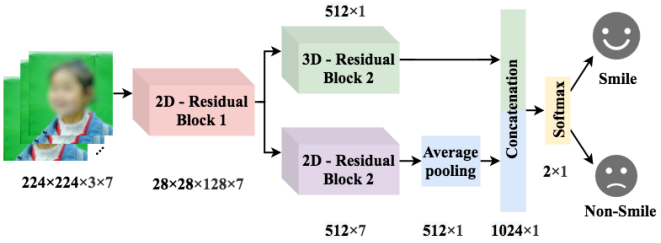

Fig. 4. Structure of the facial expression recognition neural network.

A. Temporal Stimulus Localization

The first task of the proposed framework is to locate the timestamp of each "Responsive Social Smile" stimulus. We deploy an automated speech recognition (ASR) system by the Kaldi toolkit [13]. The speech recognition model is trained from the AISHELL-2 database [24], which provides 1,000 hours of clean read-speech Mandarin data. As our ASR system can easily detect the six kinds of stimuli, then the start time of each stimulus is captured. Within the subsequent 3-second time window, the participant's response to the given stimulus will be analyzed.

B. Face Detection

The next step is to extract face images of the participant from the time-window after each stimulus. Considering young children's faces usually occupy small regions in video frames, it is necessary to find a robust face detector. Hence we choose the face detector integrated into OpenCV-DNN toolkit [25], which has been successfully tested to work well in many industrial cases. According to the detection results, a sequence of face images during the given time window could be obtained and then fed to the next step for emotion analysis.

C. Facial Expression Recognition

Analyzing children's facial expressions is a critical task for evaluating their responses to stimuli. As the facial expression is always delivered by a sequence of continuous facial muscle movements, we propose a CNN-based model to capture both spatial and temporal features for facial expression recognition. The network structure (shown in Figure 4) is mainly made of 4 parts: the 2D residual block, the 3D-CNN branch, the 2D-CNN branch, and the fusion layer. The input of our proposed model is set to a 7-frame sequence of face images that are resized to the shape of 224×224. Table IV-C illustrates the detailed architecture of the proposed model.

At the beginning of the model, a 2D residual block can extract shallow features from face images, consisting of the first three layers of ResNet-18 [26]. Moreover, the output of this block has 128 feature maps with the shape of 28×28.

In the 3D-CNN branch, extracted features from the previous block are delivered to a 3D-CNN module consisting of the fourth and fifth layers of the pre-trained 3D-ResNet-18 [27], [28]. Furthermore, the output of this part is a 512-dimensional vector. After the advent of 3D convolutional neural networks proposed in 2012 [29], this approach significantly outperforms many current state-of-the-art results in different domains.

In the 2D-CNN branch, the network is used to capture in-depth spatial information. This branch consists of the first five layers in ResNet-18 [26], followed by an average pooling layer. The output of residual parts has the shape of 512×7, while the average pooling layer will ensure that the output of this entire branch is also a 512-dimensional vector.

Finally, a fusion layer is placed to fuse the temporal and spatial information from two separate branches. At this stage, the outputs of two streams are concatenated to a 1024-dimensional vector and followed by a fully-connected output layer to predict six categories of facial expressions. We train the proposed 2D-3D CNN model on the Oulu-CASIA database [30], and Table IV demonstrates the results compared to other methods.

Since our target is to evaluate whether a child gives a responsive smile to an external stimulus, instead of classifying six kinds of emotions, the neural network trained on the Oulu-CASIA database is considered a pre-trained model. In the subsequent steps, it will be fine-tuned to a binary classifier for Smile or Non-smile.

TABLE III
ARCHITECTURE OF THE FACIAL EXPRESSION RECOGNITION MODEL

Layer Name	2D CNN Branch	3D CNN Branch
conv1	7×7, 64, stride 2	
conv2	3×3 max pool, stride 2	
	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	
conv3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
conv4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$
conv5	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$
pooling	average pooling	None
merge	concatenation, Softmax	

D. Stimulus Scoring

Each stimulus for a child is called a trial, and this section introduces how we determine the stimulus score by observing the children's facial expressions. In previous parts of this paper, we have described how to localize the corresponding time window of children's responses, the methods of extracting face images, and taking facial expression recognition. After obtaining a sequence of face images from the localized time window, we cut this sequence to several short clips with a fixed length of 7 frames. Then, each clip will be fed to the proposed facial expression classifier and get a single score to indicate whether a child gives a responsive smile. By counting the number of smiles in total clips, the stimulus score will be marked as 0 if the number is over a threshold. Otherwise, it will be marked as 1.

E. ASD Classification

Our ultimate goal is early screening for young children with ASD. Adopting all 6 stimulus scores, we concatenate each binary score to a 6-dimensional feature vector. Since this feature vector implies the child's social responses under a series of external stimuli, it is significant to be used in autism screening. Based on this feature vector, a machine learning algorithm can be trained to classify children with or without ASD.

V. EXPERIMENTS

A. Experiment Settings

As described in previous sections, each child participating in our experiment undergoes a series of stimuli from the clinician or parent. Meanwhile, the camera and microphone record the audio-visual data of the child's reactions. From the recorded data, we can locate the time window of each stimulus, extract face images, perform facial expression recognition tasks, and finally classify the ASD label.

TABLE IV
COMPARISONS OF FACIAL EXPRESSION RECOGNITION ON OULU-CASIA

Method	Descriptor	Accuracy
Yu et al. [31]	DCPN	86.23%
Jung et al. [32]	CNN-DNN	81.46%
Zhang et al. [33]	PHRNN-MSCNN	86.25%
Kuo et al. [34]	CNN	91.67%
Ours	2D-3D CNNs	89.10%

To be more specific, the default length of the localized time window is 20 seconds under the condition of 24 FPS (frames per second). Usually, there are approximately 480 frames for each stimulus. Afterward, those face images will be resized to the shape of 224×224 and cut into 7-frame clips, followed by the facial expression recognition and protocol scoring. Finally, we test several machine learning models on our clinical database using the feature vector made of wise stimulus scores. A decision tree classifier obtains the highest accuracy of 80.49%.

B. Fine-tuning FER Model

Our FER (facial expression recognition) model is initially trained on the Oulu-CASIA database [30], which includes six categories of facial expressions from 80 adults. There are two major problems that we need to solve.

- The output of the pre-trained model has six categories, which does not match with our binary classification.
- Most databases for facial expression recognition are collected from adults, which may not work well on young children.

To address these issues, firstly, we replace the output layer of the pre-trained model with a 2-dimensional one, while the trained parameters of previous layers are preserved. Besides, we collect and set up a facial expression database that is designed explicitly for young children. Fine-tuning the pre-trained model on this database improves the model's performance and usability in our experiments significantly.

The new facial expression database contains 15,000 videos (each with a length of 7 frames), and each video is manually labeled as a smile or non-smile. The videos are recorded from another 54 children under the same condition described in Figure 1. There is no overlap between these 54 children and those 41 children in our database. We fine-tune the model on our self-labeled database and finally achieve the accuracy of 92.60% for smile classification.

C. Results of Stimulus Scoring

In this part, the well-trained model of smile classification is used to give a stimulus score. As the stimulus scoring strategy introduced in our proposed framework, we employ the smile classification model to each trial and then set a threshold to determine a stimulus score for every stimulus. If the output value is higher than the limit, the related trial is marked as a smile. Otherwise, it is marked as non-smile. Since sometimes

TABLE V
CONFUSION MATRIX OF STIMULUS SCORING ON THE COLLECTED CLINICAL DATABASE

		Predicted	
		Smile	Non-Smile
Actual	Smile	62	13
	Non-Smile	16	105

TABLE VI
CONFUSION MATRIX OF ASD CLASSIFICATION BASED ON PREDICTED STIMULUS SCORES

		Predicted	
		ASD	Non-ASD
Actual	ASD	17	3
	Non-ASD	5	16

the emotion in a young child's face is ambiguous, it is difficult to distinguish the smile from other facial expressions in a toddle's face image. Therefore, we set the decision threshold to be 0.9 empirically, which means the child must give a clear enough response to count as smiling.

In our experiments, there are 41 children with or without ASD (described in III-B); each child is tested up to 6 stimuli. We finally obtain 196 stimulus scores.

To evaluate the performance of our scoring method, we invite three clinicians to watch the protocol videos and label each trial as the ground truth. Each clinician works individually, and majority voting is adopted as the final score of a given trial. Considering the limited amount of data, we employ the "leave-one-out" cross-validation strategy to evaluate the proposed method. Table V illustrates the confusion matrix of our scoring results. The experimental result shows that our method achieves an accuracy of 85.20% for predicting stimulus scores.

D. Results of ASD Classification

Based on the 6-dimensional feature vector consisting of all stimulus scores, we utilize machine learning algorithms to classify whether a child has ASD or not. As children may not always cooperate with our experiments, some missing data is set to the mean of the other stimulus scores from the same child.

Using feature vectors made of predicted stimulus scores, we evaluate four widely-used methods with a "leave-one-out" cross-validation strategy on the ASD classification task (shown in VII), and the decision tree classifier achieves the highest accuracy, sensitivity, and specificity. Table VI shows the confusion matrix of the decision tree classifier, which represents the best performance we have achieved.

To evaluate our predicted stimulus scores, we also directly utilize stimulus scores marked by the clinicians to repeat the model training process. Table VIII reveals that accurate stimulus scoring can relatively improve the performance of ASD classification. Meanwhile, it also shows that our framework performs very close to conventional clinicians.

E. Failure Case Study

To take an in-depth analysis of our proposed framework, we scrutinize the 29 stimulus scores, which are incorrectly predicted in Table V. Reviews show that most of those trials involve smiles, while doctors consider that those smiles are their spontaneous emotion rather than responses to social environments. Figure 5 demonstrates four examples of failure cases. It can be seen that these children react obviously in a smiling manner. However, the reactions are not due to designed stimuli, and they are attracted by other things nearby.

Since autistic children are heterogeneous, it is difficult to perform the ASD screening from a single aspect perfectly. The analysis above also shows that we need to fuse results from multiple protocols of the same child together to enhance the performance of the screening framework in the future.

TABLE VII
ASD CLASSIFICATION BASED ON PREDICTED STIMULUS SCORES

Algorithm	Accuracy	Sensitivity	Specificity
Logistic Regression	63.41%	66.67%	63.64%
Naive Bayes	68.29%	65.00%	68.42%
SVM	70.73%	70.00%	70.00%
Decision Tree	80.49%	85.00%	77.27%

TABLE VIII
ASD CLASSIFICATION BASED ON CLINICIAN'S STIMULUS SCORES

Algorithm	Accuracy	Sensitivity	Specificity
Logistic Regression	70.73%	70.00%	70.00%
Naive Bayes	73.17%	75.00%	71.43%
SVM	75.61%	70.00%	77.78%
Decision Tree	82.93%	80.00%	84.21%

VI. CONCLUSION

We design a standardized protocol and experiment setup for behavior analysis in ASD screening, namely "Responsive Social Smile." Also, we present a machine learning based assessment framework to predict the behavior scores for children under three years old. To improve the reliability of the proposed framework, we collect and label a facial expression database dedicated to young children, then fine-tune our facial expression module to obtain an accuracy of 92.60% on the collected emotion database. Finally, the proposed stimulus scoring and ASD classification methods obtain an accuracy of 85.20% and an accuracy of 80.49% on the clinical database.

The experiments indicate that our proposed framework can work well for ASD screening. The performance is close to clinicians' average ASD screening ability in this "Responsive Social Smile" protocol. In the future, we will fuse data from multiple complementary protocols of a child to further enhance the screening performance.

Fig. 5. Examples of failure cases.

ACKNOWLEDGMENT

The authors are grateful to people who helped with this research. Thanks to our collaborators at The Third Affiliated Hospital, Sun Yat-sen University: Yixiang Xie contributed to collect and organize the clinical database, and three clinicians helped to annotate the stimulus scores. We express our biggest thanks to the participants and their families. With their consent, this research can continue successfully.

This research is funded in part by the National Natural Science Foundation of China (61773413), Key Research and Development Program of Jiangsu Province (BE2019054), Six talent peaks project in Jiangsu Province (JY-074), Guangzhou Municipal People's Livelihood Science and Technology Plan (201903010040).

REFERENCES

[1] J. N. Constantino, C. P. Gruber, S. Davis, S. Hayes, N. Passanante, and T. Przybeck, "The factor structure of autistic traits," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 4, pp. 719–726, 2004.

[2] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of autism and developmental disorders*, vol. 37, no. 4, p. 613, 2007.

[3] W. Liu, X. Yu, B. Raj, L. Yi, X. Zou, and M. Li, "Efficient autism spectrum disorder prediction with eye movement: A machine learning framework," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 649–655.

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

[5] B. Li, A. Sharma, J. Meng, S. Purushwalkam, and E. Gowen, "Applying machine learning to identify autistic adults using imitation: An exploratory study," *PloS one*, vol. 12, no. 8, 2017.

[6] Y. Nakai, T. Takiguchi, G. Matsui, N. Yamaoka, and S. Takada, "Detecting abnormal voice prosody through single-word utterances in children with autism spectrum disorders: machine-learning-based voice analysis versus speech therapists," *Perceptual and Motor Skills*, vol. 124, no. 5, pp. 961–973, 2017.

[7] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.

[8] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[9] D. Robins, D. Fein, and M. Barton, "Modified checklist for autism in toddlers (m-chat) follow-up interview," *Publisher: Author*, 1999.

[10] 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, 2014.

[11] E. Rellini, D. Tortolani, S. Trillo, S. Carbone, and F. Montecchi, "Childhood autism rating scale (cars) and autism behavior checklist (abc) correspondence and conflicts with dsm-iv criteria in diagnosis of autism," *Journal of autism and developmental disorders*, vol. 34, no. 6, pp. 703–708, 2004.

[12] J. Kosmicki, V. Sochat, M. Duda, and D. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Translational psychiatry*, vol. 5, no. 2, p. e514, 2015.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[14] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[15] ——, "Max-margin object detection," *arXiv preprint arXiv:1502.00046*, 2015.

[16] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3267–3276.

[17] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[18] J. Hashemi, G. Dawson, K. L. Carpenter, K. Campbell, Q. Qiu, S. Espinosa, S. Marsan, J. P. Baker, H. L. Egger, and G. Sapiro, "Computer vision analysis for quantification of autism risk behaviors," *IEEE Transactions on Affective Computing*, 2018.

[19] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Computer Speech & Language*, vol. 56, pp. 80–94, 2019.

[20] W. Liu, T. Zhou, C. Zhang, X. Zou, and M. Li, "Response to name: A dataset and a multimodal machine learning framework towards autism study," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 178–183.

[21] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[22] C. C. Bell, "Dsm-iv: diagnostic and statistical manual of mental disorders," *Jama*, vol. 272, no. 10, pp. 828–829, 1994.

[23] H. E. Nag, A. Nordgren, B.-M. Anderlid, and T. Nærland, "Reversed gender ratio of autism spectrum disorder in smith-magenis syndrome," *Molecular autism*, vol. 9, no. 1, p. 1, 2018.

[24] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[25] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[28] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[30] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.
[31] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer*, vol. 34, no. 12, pp. 1691–1699, 2018.
[32] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.
[33] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.
[34] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2121–2129.