2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM) | 978-1-7281-6267-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/HEALTHCOM49281.2021.9398924

Machine Learning Based Autism Spectrum Disorder Detection from Videos

Chongruo Wu [1*] Sidrah Liaqat [2*] Halil Helvaci [2] Sen-ching Samson Cheung [2,3]
Chen-Nee Chuah [3] Sally Ozonoff [4] Gregory Young [4]

[1] *Department of Computer Science, University of California, Davis, CA, US*
[2] *Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY, US*
[3] *Department of Electrical and Computer Engineering, University of California, Davis, CA, US*
[4] *UC Davis MIND Institute, University of California, Davis, CA, US*

crwu@ucdavis.edu sidrah.liaqat@uky.edu halil.helvaci@uky.edu
sccheung@ieee.org chuah@ucdavis.edu sozonoff@ucdavis.edu gnuoysg@gmail.com

***Abstract*—Early diagnosis of Autism Spectrum Disorder (ASD) is crucial for best outcomes to interventions. In this paper, we present a machine learning (ML) approach to ASD diagnosis based on identifying specific behaviors from videos of infants of ages 6 through 36 months. The behaviors of interest include directed gaze towards faces or objects of interest, positive affect, and vocalization. The dataset consists of 2000 videos of 3-minute duration with these behaviors manually coded by expert raters. Moreover, the dataset has statistical features including duration and frequency of the above mentioned behaviors in the video collection as well as independent ASD diagnosis by clinicians. We tackle the ML problem in a two-stage approach. Firstly, we develop deep learning models for automatic identification of clinically relevant behaviors exhibited by infants in a one-on-one interaction setting with parents or expert clinicians. We report baseline results of behavior classification using two methods: (1) image based model (2) facial behavior features based model. We achieve 70% accuracy for smile, 68% accuracy for look face, 67% for look object and 53% accuracy for vocalization. Secondly, we focus on ASD diagnosis prediction by applying a feature selection process to identify the most significant statistical behavioral features and a over and under sampling process to mitigate the class imbalance, followed by developing a baseline ML classifier to achieve an accuracy of 82% for ASD diagnosis.**

***Index Terms*—Autism Spectrum Disorder, Machine Learning, Human Behavior Detection, Facial Keypoint Detection**

I. Introduction

Autism Spectrum Disorder (ASD) is one of the most prevalent developmental disorders in the U.S., affecting 1 out of 54 children according to a recent estimate [1]. Children with ASD have deficiency in social communication and show atypical attention to others and objects. For example, during social conversations, individuals with ASD tend to avoid eye contact and their eye gazes exhibit visual field scanning patterns that are significantly different from those of neurotypical individuals [2], [3]. Co-morbid conditions of ASD include epilepsy disorder, gastrointestinal disorder, sleep disruption, and others [4]. Families with ASD children are typically under significant stress, facing heavy financial burden and time commitment for their care. As such, early diagnosis and interventions of ASD are essential for both children and their families.

Previous research [5] found that early intensive interventions could have positive effects for children diagnosed with ASD. These early interventions, such as behavioral therapy and family training, help children reduce challenging behaviors and improve learning skills. Thus, accurate and widespread early diagnosis of ASD, which is a prerequisite for treatments, is important to ensure the best outcomes for children with high risks of ASD. However, early diagnosis is challenging due to the lack of access to proper medical screening and the symptoms of ASD can be missed, especially in the early years of a child. These two limitations of early screening may result in treatment delay and missed opportunity for improved treatment outcomes.

Most of the ASD screening methods depend on questionnaires or ratings [6]. They are conducted by trained professional examiners, which limits their use of these approaches in communities that lack licensed healthcare professionals and public health resources. Another drawback is that these assessments take several hours and children may have to undergo these assessments at different ages for accurate diagnoses. Hence, developing automated and efficient ASD screening tools can significantly minimize the demands on healthcare infrastructure and reduce the burden on the children. One such direction is the use of machine learning techniques, especially deep learning, in performing automated diagnosis of ASD [7]–[9].

Deep learning methods have attracted significant attention and made great contributions in various fields, such as computer vision, natural language processing, speech recognition and robotics. Instead of designing hand-crafted features, deep learning methods can automatically discover contributing features for the learning task. ASD diagnosis can be modeled as a classification problem and deep learning approaches can be

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH121344-01. Chongruo Wu was also supported by the UC Davis 2019 College of Engineering Dean's Collaborative Research (DECOR) Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

*The first two authors contributed equally to this paper.

978-1-7281-6267-6/21/$31.00 ©2021 IEEE

applied to improve prediction performance and generalization to unseen data. Previous works have explored this direction by predicting ASD based on brain imaging data [7], hand-crafted features [8] or tracked gaze data when children watch movies on a tablet [9].

Due to the large number of parameters used in a deep learning model, one of the key challenges is the availability of a large labeled dataset. This is particularly true for methods like transfer learning that can mitigate limited data but may not work for the diverse set of problems encountered in medical domains. While there are early successes in applying deep learning techniques for ASD diagnosis, most of the proposed techniques are based on small datasets and the generalization of the developed techniques are questionable. In this work, our interdisciplinary team proposed a framework that leverages deep learning and computer vision to automatically classify autism symptoms from images and videos. The technical development leverages a large video dataset that was carefully curated at the Medical Investigation of Neurodevelopmental Disorders (MIND) Institute at University of California, Davis. Motivated by the observation that children's responses in social interactions are vital to ASD diagnosis, this database focuses on recording social interaction between a child and an adult, and ASD-related behaviors were meticulously labeled at the frame level. For the machine learning perspective, we propose two pipelines to identify ASD-related behaviors in the dataset. The first one aims to apply the deep learning method to predict actions from raw images. This pipeline is motivated by the fact that deep learning approach could automatically discover helpful features which may not be observable by humans. The second approach extracts expert features including facial landmarks, head pose, and eye gaze, and then trains a deep classifier on these features to detect various behaviors. Finally, we investigate various approaches to utilize the statistics of these behaviors events in predicting ASD diagnosis.

II. Related Work

In recent years, significant work has emerged from the autism research community utilizing computer vision and machine learning in analyzing children's behaviors towards either pre-selected videos or social interactions with adults.

A. Observing subject response to pre-selected videos

The Sapiro Lab at Duke University has developed an iPhone app called ResearchKit to record subject's reaction to a video showing a social and a non-social scene displayed side by side [10]. Using this setup, it has been demonstrated that the ASD group exhibited lower likelihood to switch attention and an overall deficit of attention [11]. Automated methods were used in detecting head pose and emotion but their efficacy in identifying ASD behaviors was not explored.

Machine learning methods are also effective in discerning differences in subject response time to various stimuli including head lag when pulled to sit, delays in walking, postural differences such as slumping and stiffness, difficulty in maintaining midline position of the head. In [12], the authors demonstrated that 8% of toddlers with ASD oriented to their name with longer mean latency as compared to 63% typically developing ones. The sensitivity was 96% and specificity 38% for ASD subjects with atypical orienting. In [13], the authors investigated the level of engagement of children in response to video stimulus by identifying their expression as positive, negative or neutral with the help of facial expression recognition software.

In [14], the authors have introduced an end-to-end system for ASD classification that uses two publicly available datasets of face images, AffectNet [15] and Emotionet [16], to train a neural network to predict ASD diagnosis based on facial expressions, facial action units (AUs), valence and arousal as extracted from subject videos. The dataset consists of 88 videos which have been collected by the authors out of which 55% are of ASD subjects and 45% are of typically developing controls. They found that features from arousal, valence, expressions, and AUs gave the best performance with F1 score of 0.76, sensitivity of 76% and a specificity of 69%. A drawback is that this dataset consists only of frontal images and videos, making it difficult to use it in the wild.

B. Extraction of ASD type behaviors from subject video in an interactive environment

The approach taken by [17] is the development of a deep neural network that trains on multiple datasets to learn 3D gaze vector from third person perspective, gaze relevant scene saliency representation, and a headpose dataset for learning large face pose variations and scenario of subject having attention outside of image. The neural network outputs the gaze angle and saliency map indicating object of interest of the subject. The proposed method achieves an AUC score of 89.6% in the gaze saliency task.

Joint attention behavior is an important social skill often lacking among children with ASD. A system for characterising joint attention behavior has been developed at the Institute of Neural Computation at the University of California, San Diego using both eye-tracking and automated object detection [18]. The average sensitivity and precision of 95.9% and 97.7% have been reported. However, in the absence of comparison with state-of-the-art, this high performance cannot be placed in context.

III. Dataset Collection

This work is based on a video dataset collected under the Infant Sibling Study conducted at the UC Davis MIND institute which is an IRB approved prospective longitudinal study intended to study the onset of autism symptoms in the first 3 years of life of infants. The evaluations were conducted at 6, 12, 18, 24 and 36 months. The infant subjects comprise of a High Risk group and a Low Risk group. The High-Risk group includes infants who have an older sibling who meets ASD criteria on both ADOS (Autism Diagnostic Observation Schedule) and SCQ (Social Communication Questionairre).

More details of the inclusion/exclusion criteria of the study can be found in [19].

The video dataset consists of 1707 videos of 365 infants engaged in various adult-child play tasks. Each video is a recording of a 3 minute interactive play session either with a parent or an examiner. As the symptoms of ASD could be reflected by the social reactions, like the eye contact, play behaviors and communication with others, our experiments are designed to monitor the reactions of children in their interactions with examiners. These interactive experiments are held in a room with cameras to record the actions of examiners and children. Sitting behind a private glass window, researchers are able to view all activities in the room, and control the recording progress. During the experiments, an examiner, who has been trained to evaluate autistic behavior, is sitting in front of a child. The parent of the child may sit at the corner in the same room. To attract the child's attention, the examiner may hold different toys, like little doll or small cellphone, and interact with the children. Examiner would also interrupt the child with another toy to test whether the child has any response to different stimulants. All performance of the child is recorded. In each video, there is a camera to monitor children's action, and another one to record the adult. Up to three play sessions can be recorded for each visit.

For each video, we focus on four different actions 1) Look face: whether the subject looks at other's face (making eye contact with the partner); 2) Look object: whether the subject looks at object that is of interest to him/her; 3) Smile: whether the subject is smiling or not; 4) Vocal: whether the subject makes sound(speaking). Examples of these four actions are shown in Fig 1. To provide ground-truths for our supervised learning tasks, these actions were manually labeled or coded by trained coders. All videos were coded in Observer 5.0, a behavioral observation software by Noldus, by coders trained to at least 90% agreement on all codes. Further, double coding by master coders on 15% of data was done to ensure reliability resulting in very good intraclass correlation coefficients (ICC) for all codes (gaze to face: 0.95, gaze to object: 0.98, smile: 0.96 and vocalization: 0.93).

Altogether, there are 1707 videos carefully annotated with these behavior episodes. In the early investigation reported in this paper, we use only 547 videos of 133 subjects to conduct experiments. In the video dataset, we have behavior labels available at the frame level resulting in approximately 2.95M examples for the behavior detection task. These videos are split into three folds: a training set, a validation set, and a test set. To verify the generalization of the trained models, there is no overlap of human subjects among these three sets. The statistics are listed in Table I.

(a) Look Face (b) Look Object

(c) Smile (d) Vocal

Fig. 1. Examples of four behaviors of interest in our dataset

TABLE I
THE STATISTICS OF TRAIN, VALIDATION AND TEST DATASETS, INCLUDING THE NUMBER OF CHILDREN, NUMBER OF VIDEOS, AND THEIR CORRESPONDING PERCENTAGE.

	#Children	#Videos	Video Percentage
Train	100	403	73.6%
Val	10	48	8.7%
Test	23	96	17.55%

IV. Behavior Detection based on Image-based Deep Learning

Directly using the current video dataset to predict ASD diagnosis is challenging. First, we only have a very limited number of video instances, especially for ASD subjects, and they are likely to be insufficient to train deep neural networks and often result in overfitting. Second, while we only have a limited number of video instances, we do have a very rich set of fine-grained manual behavior labels for each video that have been shown to be useful in distinguishing ASD from TD [19]. As such, leveraging intermediate supervised signals could ease the network learning.

Based on the two reasons above, in the initial stage, our goal is to obtain accurate predictions of different actions in each video clip, which we will then utilize to infer the final diagnosis. We treat the problem as image-based classification task. Deep neural network is trained on all frames. With the same artechitecture, we train different models in an end-to-end fashion for different action events. In this Section, we consider the approach of end-to-end training to let the deep neural network determine the optimal image features to identify each action.

Binary cross entropy is used as the loss function. The model is trained by using SGD optimizer. Learning rate is initialized as 0.01 and will be decreased by 10x every 10 epochs. We use ResNet-18 as our backbone and it is pretrained on the ImageNet dataset. We resize each image frame to 224x224 in order to be compatible with pretrained ResNet. For the data augmentation, we horizontally flip the input image during the training.

Our dataset is not balanced as the number of negative

samples for each action is on average fives times more than the number positive samples. During the training, larger negative samples may dominate training process, leading to a bias towards the majority class. Inspired by Holistically-Nested Edge Detection [20], we apply the weighted loss to mitigate this issue during the training. The final loss is

$$\alpha \, \Sigma_{Y_+} \log P(y = 1|X, W) + \beta \, \Sigma_{Y_-} \log P(y = 0|X, W)$$

Where $\alpha = |Y_-|/|Y|$ and $\beta = |Y_+|/|Y|$. $|Y_+|$ and $|Y_-|$ denote the number of samples for positive and negative samples. X and W represent the input image and network weight. With this loss, the penalties for larger negative samples become smaller and it makes network pay less attention on them.

V. Behavior Detection based on Facial Keypoints as Features

An alternative to the image based end-to-end approach in Section IV is to train a neural network on well-established facial features extracted from the child's face in a supervised manner. These features include face and eye landmarks, facial action units (AUs), head pose, and eye gaze direction, extracted by the OpenFace 2.0 library [21]. The model is trained independently for three behaviors namely: gaze to face, smile and vocalization in a binary classification setting using frame level binary behavior labels derived from the expert annotations. As only facial features are used, the classifier for look object is not built. Fig 2 shows a screenshot of our visualization tool with facial landmarks on top and a wrap-around timeline at the bottom depicting the ground-truth intervals in blue, predicted intervals in red, and current position as a scrolling yellow line.

To leverage the information coming from different modalities, we used a hybrid architecture with a 1D convolution branch with three layers and a fully connected branch with four layers. The input to the 1D convolution channel is facial landmarks which include 56 eye region landmarks and 68 face landmarks. The inputs to the dense channel are: rotation and translation parameters of child's head pose, 2D gaze parameters and 17 FAU features along with the confidence score of FAUs. The output of these two channels are then flattened and combined through a small dense network with four fully connected layers. The model is trained using Adam optimizer with an initial learning rate of 1e-3 on standard binary entropy loss. Class weighting on the loss function is found to give the best performance for all three tasks.

VI. From Behaviors to ASD Diagnosis

In Sections IV and V, we have proposed automated techniques to detect the behavioral events of look face, look object, vocalization, and smile. In [19], these events alongside with the created variables of shared smile, when the smile code overlapped with the look face code, and shared vocalizations, when a child vocalized at the same time that he/she was looking at the adult, were shown to be significant when diagnosing ASD. Specifically, the authors found that there is a statistically significant group difference in the aggregate

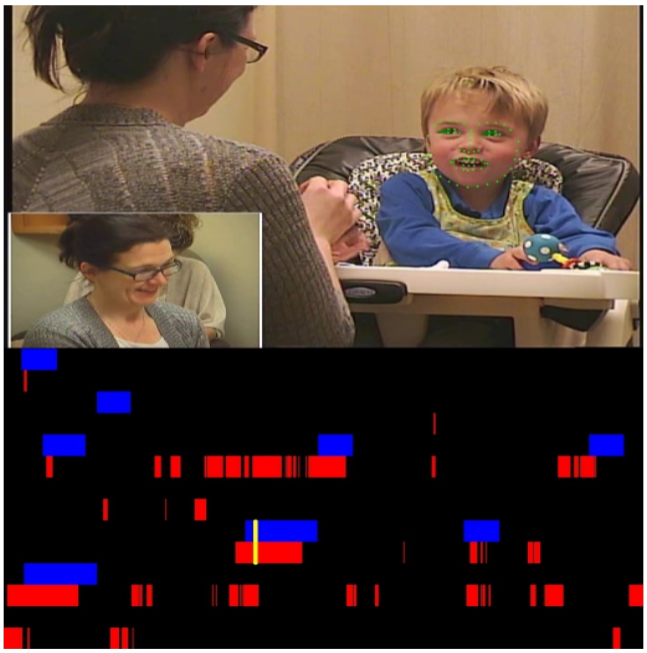

Fig. 2. Video frames with facial keypoints and gaze along with snapshot of ground truth (blue) vs predicted labels (red) for look face on a test video

statistics of these events between the ASD and TD subject groups. In this section, we cast this as a machine learning problem and study the appropriate algorithms for feature selection, class rebalancing and classifier construction. In our study, we use 22 features including the total occurrences, total duration, rate (number of occurrences over length of video), and proportion (total duration over length of video) of the six aforementioned events, the gender of the subject, the age, the risk in our feature set over a set of 1707 videos, a superset of the video dataset used in Sections IV and V. The outcome variable is the ASD diagnosis conducted at the time of the visit by a clinician when the video recording was made.

Feature selection techniques are applied to the dataset to select relevant variables and also eliminate highly-correlated variables. To have an initial understanding of how different features are related to one another, a correlation matrix for all the features were calculated. The results showed that the occurrences, duration and rate values for each event are highly correlated. To avoid redundant use of related variables, only rate values are included in the study. A number of different statistical methods are applied to reveal the most important characteristics that contribute to ASD diagnosis. The techniques applied in this study are: Recursive Feature Elimination (RFE), Ridge Regression (RR), Mutual Information Estimation (MI) and Kolmogorov Smirnov Test (KS) [22]. The motivation of choosing RFE and RR is that they are commonly used techniques in literature. However, they are restrictive in their application because they assume the features follow a normal distribution with equal variances. On the other hand, MI and KS are non-parametric, meaning that they do not assume any underlying distribution to the dataset. Both parametric and

non-parametric techniques are chosen to verify that the most important features have been correctly selected. The results in Section VII show that the majority of these techniques also the rate variables to be the most significant.

To evaluate the feature selection performance, two independent neural networks (NN) are constructed based on the full and selected sets of features. Both models consist of 3 fully connected layers. The Scaled Exponential Linear Unit (SELU) activation function is used on the first two layers and the Softmax function is used in the output layer. The first hidden layer has 128 nodes, the second hidden layer has 256 nodes and the output layer has 2 nodes. Cross entropy is used as the loss function and Adam for optimization. The first network uses inputs of look face rate, social smile rate, social vocal rate, age and gender, which are the more discriminating features identified earlier. The second network is trained using all the variables in the dataset. Similar to the training of the behavior detection classifiers, all videos of the same subjects are either in the training or testing set but not both.

There is a significant class imbalance between the ASD and Non-ASD samples in the dataset. 93.9% of the combined dataset belongs to the Non-ASD class. Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links undersampling method [23] are two commonly used techniques in alleviating the problem of class imbalance. Using the complete feature set from the dataset, the effects of oversampling and undersampling on class imbalance have been evaluated for the same network architecture. In the first network the training set is rebalanced using SMOTE only. For the second network only Tomek Links is used to resample the training set. For the third network, both SMOTE and Tomek Links are applied to rebalance the training set.

VII. Experimental Results

A. Results of image-based method

Table II indicates the performance of the image based method on detecting the four behaviors. Given the total numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the following metrics are commonly used in the literature: accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$, sensitivity or recall (RE) $= \frac{TP}{FN+FP}$, specificity $= \frac{TN}{FP+TN}$, precision (PR) $= \frac{TP}{TP+FP}$, and F1 score $= \frac{2PR \cdot RE}{PR+RE}$. Accuracy is typically used in ML literature, while sensitivity and specificity are used in medical literature. F1-score provides a single number to summarize the overall performance. As such, these four metrics are used in this paper. As the numbers of positive and negative categories are imbalanced in our test dataset, we use macro-average which weighs positive and negative classes equally for a balanced results.

In the table, we can see the model performs quite well on three actions: smile, look face and look object. The F1-score and macro-average are both greater than 67%. However, the performance on detecting vocalization is significantly worse because no audio information is used in the detection. The combination of image and audio as a unified input to a deep network is being investigated.

TABLE II
Results of image-based method

	Sensitivity	Specificity	F1-score	Accuracy
Smile	0.47	0.93	0.72	70%
Look Face	0.39	0.98	0.72	68%
Look object	0.97	0.36	0.71	67%
Vocal	0.18	0.89	0.53	53%

B. Results of Facial keypoints based method

Table III shows the baseline performance for smile, look face and vocalization detection based on facial behavior features of the child subject. These results reflect the aggregate performance on a held-out test set of videos where there is no overlap of subjects in the training and test data. For the look face task, a sensitivity of 0.59 and specificity of 0.73 indicate that the model is performing well based on the child's features alone. It is reasonable to assume that incorporating information about the head position and facial features of the interacting adult would lead to further improvements in the look face task. For smile detection, specificity of 0.92 is quite good however sensitivity of 0.45 has room for improvement. A study of emotion detection literature reveals facial landmarks and FAUs to be very useful features especially for smile detection when the face is in frontal or near frontal position. The performance on this dataset is hampered by the fact that the child's face is often in a non-frontal position due to the camera position. Moreover, face occlusions by toy objects or head movement frequently occur and lead to failed face detections. The low sensitivity for vocal task indicates that this model is not well suited for detection of vocalization task since a frame based approach cannot make use of the motion of lips. A temporal processing approach might be useful to improve the detection of the mouth movement.

TABLE III
Results of Facial keypoints based method

	Sensitivity	Specificity	F1-score	Accuracy
Smile	0.45	0.92	0.71	68.5%
Look Face	0.59	0.73	0.60	66.0%
Vocal	0.06	0.94	0.48	50.0%

C. Results of behaviors to ASD diagnosis

The events in the dataset are divided into two groups: Video-related and non-video related. The first group contains all statistical features related to the behaviors related to events that occur in the video. The second group is the background information of the child subject. In order to eliminate error introduced by the automated techniques, the first group of features are derived directly from the ground-truth data.

The three most important video related features determined by the MI, RR and KS methods are look face rate, social smile rate and social vocal rate. The RFE method identifies look face rate, vocal rate and smile rate to be the most important features. Non-video related events determined as necessary by all of the methods are age and gender. Since MI, RR and KS

all agree to have the same top video related events, results obtained from RFE are ignored. These findings are similar to those determined in our earlier work [19].

TABLE IV
NEURAL NETWORK RESULTS

	Sensitivity	Specificity	F1 Score	Accuracy
All Features	0.04	1.00	0.52	58%
Selected Features	0.02	1.00	0.50	51%
SMOTE	0.83	0.70	0.55	77%
TomekLinks	0.08	1.00	0.56	54%
SMOTE + TomekLinks	0.92	0.71	0.58	82%

Table IV shows the classification performance on the test dataset using different feature sets and class rebalancing approachs. Performance using only the selected features is similar to the performance achieved using all features. Applying SMOTE to the training set significantly increased the sensitivity while reduced the specificity. This shows that the synthetic features generated by SMOTE can improve the detection of the ASD class to a certain extent. Implementing Tomek Links slightly improved the performance of the network, indicating that the links removed from the negative class did not necessarily solve the class imbalance issue. Applying SMOTE then Tomek Links significantly improved the sensitivity of the model and slightly improved the specificity of the model over SMOTE only. The combined re-balancing scheme produces the overall best results based on the F1 scores and accuracy.

VIII. Conclusions

In this paper, we have introduced a large-scale video dataset of children's social interaction with manual labeling of behaviors that are clinically significant for ASD diagnosis. A machine learning framework has been proposed for this dataset. This framework consists of two stages: behavior detection and ASD prediction based on statistical features of behaviors. For behavior detection, we have proposed two baseline deep-learning techniques, one based on end-to-end training on raw video frames while the other relied on facial features of the child subject. For ASD prediction, we have investigated feature selection, class rebalancing, and neural network classifiers to link the statistics of behaviors to ASD diagnosis. In the future, we plan to incorporate temporal dimension for general improvement in accuracy of behavior detection, utilization of audio information for vocalization detection, and self-supervised schemes to detect adult's head and objects of interest in image frames to improve look-face and look-object detection.

References

[1] *Data Statistics on Autism Spectrum Disorder*. [Online]. Available: https://www.cdc.gov/ncbddd/autism/data.html

[2] L. Ronconi, M. Devita, M. Molteni, S. Gori, and A. Facoetti, "Brief report: When large becomes slow: Zooming-out visual attention is associated to orienting deficits in autism," *Journal of autism and developmental disorders*, vol. 48, no. 7, pp. 2577–2584, 2018.

[3] W. Jones and A. Klin, "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, 2013.

[4] *Data Statistics on Autism Spectrum Disorder*. [Online]. Available: https://www.chop.edu/news/autism-s-clinical-companions-frequent-comorbidities-asd

[5] S. Camarata, "Early identification and early intervention in autism spectrum disorders: Accurate and effective?" *International Journal of Speech-Language Pathology*, vol. 16, no. 1, pp. 1–10, 2014.

[6] F. Thabtah and D. Peebles, "Early autism screening: A comprehensive review," *International journal of environmental research and public health*, vol. 16, no. 18, p. 3502, 2019.

[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] I. M. Nasser, M. Al-Shawwa, and S. S. Abu-Naser, "Artificial neural network for diagnose autism spectrum disorder," 2019.

[9] G. Dawson, K. Campbell, J. Hashemi, S. J. Lippmann, V. Smith, K. Carpenter, H. Egger, S. Espinosa, S. Vermeer, J. Baker *et al.*, "Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder," *Scientific reports*, vol. 8, no. 1, pp. 1–7, 2018.

[10] G. Sapiro, J. Hashemi, and G. Dawson, "Computer vision and behavioral phenotyping: an autism case study," *Current Opinion in Biomedical Engineering*, vol. 9, pp. 14–20, 2019.

[11] M. D. M. J. Bovery, G. Dawson, J. Hashemi, and G. Sapiro, "A scalable off-the-shelf framework for measuring patterns of attention in young children and its application in autism spectrum disorder," *IEEE Transactions on Affective Computing*, 2019.

[12] K. Campbell, K. L. Carpenter, J. Hashemi, S. Espinosa, S. Marsan, J. S. Borg, Z. Chang, Q. Qiu, S. Vermeer, E. Adler *et al.*, "Computer vision analysis captures atypical attention in toddlers with autism," *Autism*, vol. 23, no. 3, pp. 619–628, 2019.

[13] 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.

[14] B. Li, S. Mehta, D. Aneja, C. Foster, P. Ventola, F. Shic, and L. Shapiro, "A facial affect analysis system for autism spectrum disorder," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4549–4553.

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[16] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[17] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 383–398.

[18] P. Venuprasad, T. Dobhal, A. Paul, T. N. Nguyen, A. Gilman, P. Cosman, and L. Chukoskie, "Characterizing joint attention behavior during real world interactions using automated object and gaze detection," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–8.

[19] S. Ozonoff, A.-M. Iosif, F. Baguio, I. C. Cook, M. M. Hill, T. Hutman, S. J. Rogers, A. Rozga, S. Sangha, M. Sigman *et al.*, "A prospective study of the emergence of early behavioral signs of autism," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 49, no. 3, pp. 256–266, 2010.

[20] S. "Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[22] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2008, vol. 207.

[23] G. E. Batista, A. L. Bazzan, and M. C. Monard, "Balancing training data for automated annotation of keywords: a case study." in *WOB*, 2003, pp. 10–18.