CHAPTER

10

Behavioral phenotype features of autism

Huiyu Duan[1], *Jesús Gutiérrez*[2], *Zhaohui Che*[1], *Patrick Le Callet*[3] *and Guangtao Zhai*[1]

[1]Shanghai Jiao Tong University, China [2]Universidad Politécnica de Madrid, Spain
[3]Nantes Université, France

10.1 Introduction

Autism spectrum disorder (ASD; henceforth "autism") is one type of neurodevelopmental condition, whose underlying neurobiology marker is still unclear. Individuals with autism show delays in the development of human cognition [1], leading to the difficulties with cognitive empathy across the lifespan [2]. The phenotype markers including social communication symptoms, fixated or restricted behaviors or interests, hyper- or hypo-sensitivity to sensory stimuli, and associated features have historically been the primary markers in the diagnosis procedure of autism [3]. However, current diagnosis procedures are time and labor expensive and require well-trained clinicians to administer, resulting in long waiting times for at-risk individuals. In this paper, we discuss several state-of-the-art techniques exploring objective and quantitative behavior phenotype features for autism, which includes atypical visual attention, action and drawing behavior. These techniques may shed light on future studies and instruments related to the analysis and computer-aided diagnosis for autism based on behavioral phenotype.

Sensory symptoms [4] are widely observed among individuals with autism [5] and may affect every modality including vision, audition, touch, smell, and taste. Atypical sensory perceptions are identified to be one of the core characteristics of autism [4]. As an important aspect of sensory perception, atypical visual attention is often observed in individuals with autism [6]. Eye movement is a representative sensory phenotype since it encodes rich information that can reflect human cognition, attention, psychological factors, and so on. Visual attention differences between individuals with autism and typically developing (TD) groups have been widely observed and frequently reported in the literature [7]. These differences include reduced joint attention behaviors [8], reduced attention to social scenes [9], and preference to low-level features of the stimuli [10], etc. In brief, individuals with autism show reduced attention to social stimuli (i.e., faces, conversations, etc.) but pay more attention to nonsocial stimuli (i.e., vehicles, electronics, etc.).

Since eye movements encode so much underlying human cognition information and many studies have demonstrated the differences in eye movements between individuals

Neural Engineering Techniques for Autism Spectrum Disorder, Volume 2
DOI: https://doi.org/10.1016/B978-0-12-824421-0.00018-7

with autism and healthy controls, recently, some researchers are exploring the use of eye movements as phenotype markers to aid the diagnosis of autism. However, the stimuli that can comprehensively and accurately differentiate two groups through eye movements still need to be explored. In this paper, we consider three types of stimuli including natural stimuli, face stimuli, gaze-following stimuli, and discuss the gaze pattern differences between autistic people and TD people on these stimuli.

Several studies have explored the visual attention of individuals with autism on natural stimuli [7,10,11]. Wang et al. conducted data-driven analysis of gaze patterns during natural scene-viewing in individuals with autism which shows the visual attention in autism are biased to low-level pixel features (such as contrast, color, and orientation) [11]. Duan et al. conducted experiments on a wide range of natural stimuli and analyzed several differential features between children with autism and TD controls [7]. Based on these differences, some deep neural network (DNN)-based studies have been performed [11,12], which tried to predict the gaze patterns of children with autism and typical controls, respectively. Moreover, a successful challenge [13] held on IEEE International Conference on Multimedia and Expo (ICME) also attracted many competitors and produced several possible solutions for both the prediction and the classification of the gaze patterns of individuals with autism and typical controls.

Since faces are important social cues, research related to the visual attention of individuals with autism on face stimuli has also attracted the attention of many researchers. Compared to the normal population, individuals with autism have impairments in face recognition or discrimination [14]. Existing eye-tracking experiments consistently demonstrate that people with autism have reduced visual attention to faces compared to the controls [15]. Regarding visual attention on core facial features, some studies found reduced visual attention of people with autism to these regions [16], while other studies reported no differences in gaze patterns between autism and typically developing individuals [17]. Moreover, the influence of facial emotions on the visual attention between autistic individuals and TD people is different [18]. Duan et al. [19] conducted experiments on face images in the wild and analyzed several features related to facial images. A DNN-based model for predicting the gaze pattern of individuals with autism on face images has also been proposed.

Another type of special social stimuli that may arise the visual attention differences between individuals with autism and TD controls is joint attention. Gaze cues can provide not only information about the locations of gaze-at objects but also complicated insights for social cognition. People can pay more attention to the object being looked at for social referential understanding, such as visual perspective-taking and empathy [20]. Falck-Ytter et al. [21] tested the spontaneous gaze and point-gesture following in autistic children. The gaze performance results showed that gaze following is closely related to adaptive communication skills and indicated joint attention impairments in autism. In another study [22], TD children showed longer first fixations to the target during the congruent condition (*i.e.*, the head of the model is oriented to and gazes at the target), while autistic children showed shorter duration. This experiment result indicated the loss of ability to enhance social salience of the gaze-at target in autistic children. Fang et al. [23] conducted the first large-scale experiment among autistic children in terms of the visual attention on gaze-following stimuli. A DNN-based classification model was also proposed to classify the autistic children and TD controls based on their gaze pattern on gaze-following stimuli, which achieved better results.

There are many other visual stimuli that may be helpful in distinguishing individuals with autism and TD people such as biological motion stimuli [24,25], etc. The atypical gaze patterns of individuals with autism still need to be further explored.

Besides eye movements, atypical action behavior is also common among autistic individuals. Behavior action phenotypes, such as fixated or restricted behaviors or interests, have historically been adopted as one important index for the diagnosis of autism [3]. However, current action phenotype diagnostic procedure still needs clinicians to roughly estimate the symptoms. To the best of our knowledge, there are no existing automatic methods that can routinely screen autism and give level inference, accordingly based on their atypical action. Objective methods or instruments are also lacking. With the rise of DNN, impressive progress has been made in action detection and recognition, which also promotes some studies on exploring automatic detection and recognition of atypical action behavior [26]. These automatic methods can help detect autism effectively and with low cost.

Finally, another atypical behavior that has not been widely studied, that is, atypical drawing behavior, may also help automatically screen autism. Art therapy for autism has been studied a lot [27] and researchers believe that drawing is a nonverbal way of communication that may help children with autism speak up. Shi et al. [28] explored the differences between the paintings drawn by autistic children and TD controls. Since drawing can also reflect the human cognition, and psychological factors, it may also reveal the hallmarks of autism. This drawing phenotype may provide a large-scale screening method for autism.

Overall, in this paper, we discuss several promising artificial intelligence (AI)-based methods to automatically and objectively characterize and detect the behavior phenotype of ASD. These methods can be further used to assist in the screening and diagnosis of autism, thereby improving the efficiency and quality of the diagnosis of autism.

10.2 Eye movement behavior phenotype of autism

In this section, we mainly focus on the phenotype of eye movement behavior. As mentioned above, individuals with autism show reduced attention to social contents. Since eye movement needs stimuli to activate, it is important to explore the gaze pattern of autistic individuals on different stimuli and design effective stimuli and procedures to aid the diagnosis of autism. Here we discuss three types of stimuli including natural stimuli, face stimuli, and gaze-following stimuli as follows. The general procedure from data collection to processing and analysis is illustrated in Fig. 10.1.

10.2.1 Natural stimuli

10.2.1.1 Dataset

Duan et al. [7] established a dataset of eye movements for children with autism on natural stimuli, which is the first large-scale open source eye movement dataset for autism. A total of 500 images selected from MIT eye-tracking dataset [29] were included in the

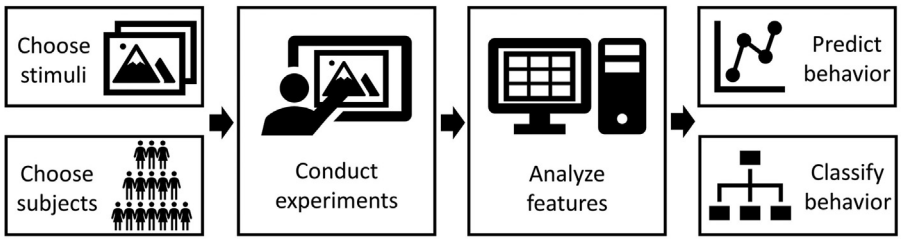


FIGURE 10.1 General procedure from data collection to data processing and analysis.

dataset. These images covered a wide range of natural scenes including animals, buildings or objects, natural scenes, one person, several people, people with objects, etc.

Twenty high-functioning autistic children were recruited in this experiment. All subjects met DSM-V diagnostic criteria for autism [3]. Since it is hard to make children with autism concentrate on the screen, only fourteen subjects could complete the calibration step and provide valid eye movement data. The ages of the remaining autistic children were between 5 and 12 years, with an average of 8 years. Fourteen typically developing controls were also recruited. These controls matched in terms of age, gender, race, education, and intelligence quotient (IQ) with the children with autism, in which IQ was assessed using the Wechsler Abbreviated Scale of Intelligence (WASI). All subjects including autistic children and controls were confirmed to have normal or corrected-to-normal vision.

During experiments, all subjects were seated at around 65 cm from the eye tracker Tobii T120, which displayed the stimuli (natural images) on a 17 inches screen. The resolution of the screen is 1280×1024. The sampling rate of the eye tracker was set to 120 Hz. To avoid eye fatigue and inattention, the experiment was divided into several short sessions. For each session, all subjects freely watched 30 images that were presented in different random orders. Each image was displayed for 3 seconds and a 1-second gray-background image was displayed between stimuli. The calibration of the eye tracker was done before each session for each participant.

10.2.1.2 Analysis

The eye movement data can be obtained from the above procedure. It is important to perform qualitative analysis based on the dataset to shed light on subsequent quantitative analysis and algorithm designing. Fig. 10.2 shows the examples of the comparison between the visual attention of autistic individuals and TD controls. The first three subfigures in the first row show a series of social activities. Fig. 10.2A demonstrate

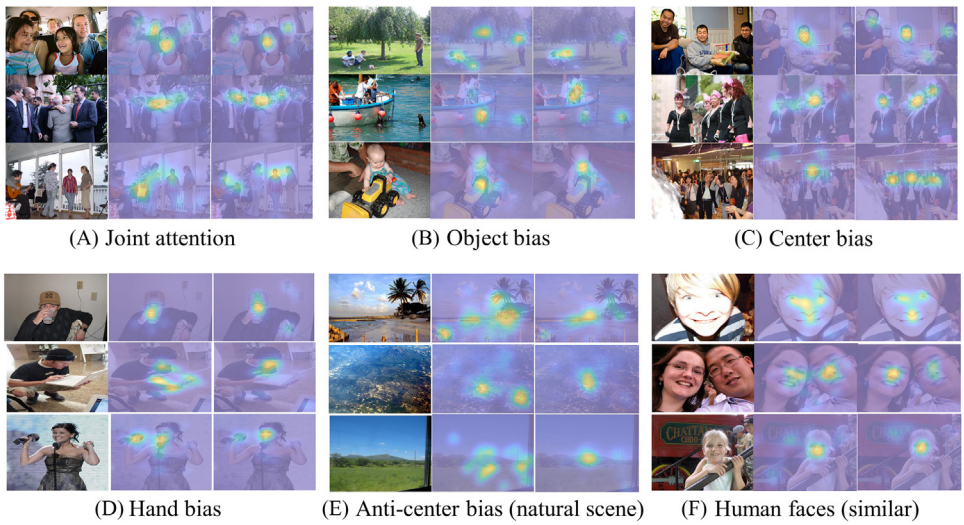


FIGURE 10.2 The visual attention comparison between children with autism and TD controls. Three columns of each subfigure from the left to the right are stimulated images, heat map of the visual attention of children with autism and heat map of the visual attention of TD controls.

the joint attention difference between children with autism and controls, where autistic children tend to focus on the people or objects of interest in the center region without the consideration of joint attention. However, as shown in Fig. 10.2B, this absence of joint attention of autism will disappear when the target is an object or an animal, and they even pay more attention to these areas of images compared with TD controls. We claim this phenomenon as objects or animals bias. The center bias as mentioned in [10] can also be widely observed in this dataset, which may reveal the unconscious attention behavior of autism.

Another observation is hand bias, which can be concluded from this dataset as demonstrated in Fig. 10.2D. In the situation where there is an interactive activity between objects and the hand of the main character, children with autism tend to focus more on the hand or the objects in the hand, while TD controls pay more attention on the face of the main character. Fig. 10.2E shows the visual attention comparison between autistic children and TD controls on landscape images. An interesting phenomenon against the center bias as described in [10] can be observed in this situation. Autistic children tend to fixate more on pixel-level salient features and the distribution of their attention on landscape images are scattered, while TD controls show obvious center bias on this kind of stimuli. Thus, the center bias of visual attention may be related to whether subjects are interested in the image, and if they do not have interest in the image, more obvious center bias can be observed. As demonstrated in Fig. 10.2F, when faces are the main contents of one image, the visual attention maps of children with autism and TD controls are similar. This is a little bit different with previous works, thus more specific discussion on face stimuli will be performed later.

The differences also exist in gaze sequence (i.e., scan path), readers can refer to [7] for more details.

10.2.1.3 Gaze pattern classification and saliency prediction

Studying the differences between visual attention patterns of individuals with ASD can be helpful in the diagnosis of this disorder and it can also help on the development of adequate tools that can improve their quality of life.

In this sense, the recognition of characteristic features in the way that individuals with ASD explore visual stimuli by analyzing their sequences of fixations can provide insights to identify them. Thus, the development of automatic models able to detect these characteristic patters can be a great support for clinicians in the diagnosis and assessment of ASD.

In addition to the visual attention patterns related to the fixation trajectory, identifying systematic behaviors of individuals with ASD related to the regions or contents within the visual stimuli can be helpful to develop computer-human interfaces specifically designed for the needs and comfort of people with ASD. In this effort, saliency models that are able to predict those regions or contents that are more important for individuals with ASD can be of great help.

In particular, these two cases (depicted in Fig. 10.3) were considered for the two tracks proposed within the Grand Challenge "Saliency4ASD: Visual attention modeling for Autism Spectrum Disorder," held at IEEE ICME '19 [13]. For both cases, images were used as visual stimuli to train and test the models [7], and only high-functioning children with ASD and children with TD were considered, without accounting neither for the possible presence of comorbidities (i.e., when two or more disorders co-occur in the same subject, such as Attention Deficit Hyperactivity Disorder), nor for different severities of ASD within the autism spectrum.

10.2.1.4 Models submitted to Saliency4ASD

For the prediction of the saliency maps that fit the gaze behavior of children with ASD, four models from four different teams were

FIGURE 10.3 Diagram of the two types of models considered in the Grand Challenge "Saliency4ASD." (A) Saliency prediction models that fit gaze behavior of people with ASD, (B) models able to classify ASD and TD viewers using gaze data.

submitted to the Grand Challenge. Their main characteristics are:

- Shanghai University & University of Rennes (SU&UR) [30]: The proposed model was based on a Dilated Convolutional Network (DCN) grounded on VGG-16 [31] with deconvolutional operations. In addition, the approach considered the fusion of multilevel features, deep supervision, and single-side clipping.
- University of Rennes & Shanghai University (UR&SU) [32]: This approach was based on a Deep Convolutional Neural Network (CNN) based on a 2-stream VGG-16 [31] and coarse-to-fine architecture. Another particularity of the model was the loss function embedded on a regularization term.
- Jiangxi University of Finance and Economics (JUFE) [33,34]: This model employed a DNN based on U-Net [35] and a novel loss function called Positive and Negative Equilibrium Mean Squared-Error (PN-MSE).
- Indian Institute of Technology Jodhpur (IITJ): This team submitted a model whose implementation details were not revealed.

In addition, a baseline model was used to obtain an idea of the added value provided by proposed models with respect to a state-of-the-art model. Thus, the SalGAN [36] model was considered without any re-training nor tuning as baseline.

Regarding the classification of an observer with ASD or TD given an image and the sequence of fixations, five different teams submitted a total of eleven models, whose main properties are:

- Technical University of Munich (TUM) [37]: This team proposed a random forest classifier using features from: scanpaths, saliency (computed with SAM-ResNet [38]) and image content (using a CNN-based face detector).
- Roma Tre University (R3U) [39,40]: The proposed model was a TreeBagger classifier. based on random forest, using features from: image content (using YOLO object detector [41]), saliency (based on SDSP [42]), fixations, and center bias.
- University of Miami (UM) [43]: A model based on CNN and Long Short-Term Memory (LSTM) architectures was proposed. Firstly, SalGAN [36] model was used to estimate image saliency. In total, this team proposed three different models

combining two variants of CNN structures and with/without batch normalization.
- Univ. of California Davis & Univ. of Kentucky (UCD&UK) [44,45]: This team submitted three models: one using an architecture with two branches (based on Resnet 46) to extract image features and to process data points models, and two models based on a fully connected dense network (FCN) trained on real and synthetic (using STAR-FC [46]) scanpaths, one using a small set of high-level features and the other using all available information. A third model was proposed.
- East China Normal University (ECNU) [47]: This approach was based on a simple classifier that analyzes the gaze-deviation distance (using a nonparametric visual model to obtain image saliency) and gaze duration time. With different settings of the parameters, three versions of the model were submitted.

In this case, the following baselines were used: a random classifier, a classifier that labeled all inputs as TD, and another one labeling all as ASD.

10.2.1.5 Evaluation criteria

To evaluate the performance of the models, the outputs provided by them for the input test data were compared with the ground-truth data (which is not known by the model developers for the sake of a fair comparison among models).

In particular, for the models that estimate the saliency of children with ASD given an input image, the following metrics were used to compare the output saliency models with from the ground truth: Normalized Scanpath Saliency (NSS), Kullback–Leibler divergence (KL), Correlation Coefficient (CC), Similarity (SIM), and Area under the curve (AUC) [48][1].

Furthermore, to evaluate the performance of the models that, given an image and a sequence of fixations from one observer, try to classify the subject between ASD or TD, typical classification metrics were used, taking into account the labels ($ASD = 1$ and $TD = 0$) provided by the models and the labels from the ground truth. In particular, the following set of metrics was used [49]: accuracy, precision, recall/sensitivity, specificity, F1-score, AUC, and Cohen's Kappa. Given the limitations of some of these metrics, this complete set was used to better compare the models. Nevertheless, a more appropriate selection of metrics will be another focus of the future research taking into account that it depends on the balance of the dataset and the specific scenario. For instance, when classifying subjects with "ASD" and "TD," a certain trade-off between high classification performance and low false negative rate (classify as "TD" subjects with "ASD") would be important.

10.2.1.6 Results of Saliency4ASD

The results of the two types of models submitted to the Grand Challenge "Saliency4ASD" according to the aforementioned set of metrics are reported in Table 10.1 and 10.2.

As it can be seen, and taking into account the properties of the models described in the previous Section 10.2.1.4 (and more details can be found in their respective papers), machine learning approaches generally provide better performance. In addition, the combination of image content features (e.g., faces, eyes, mouth, presence and size of other objects, contextual semantics, etc.) and features related to exploration biases of people with ASD (e.g., fixations close to the center of the image, fixation durations, etc.) provide significant advantages in contrast with other features [13]. Finally, as reflected by the reported numbers for the performance metrics, it is worth noting that further research is needed to obtain better performing models.

[1] Available at http://saliency.mit.edu/

TABLE 10.1 Performance for the set of saliency metrics of the submitted models for saliency prediction of individuals with ASD.

Team	NSS (↑)	CC (↑)	SIM (↑)	KL (↓)	AUC_Judd (↑)	AUC_Borji (↑)	Rank
SU&UR [30]	1.510	0.681	0.623	0.590	0.818	0.786	1
UR&SU [32]	1.419	0.682	0.631	0.902	0.811	0.785	2
JUFE [33,34]	1.245	0.600	0.587	0.632	0.790	0.769	3
IITJ	0.656	0.316	0.468	0.911	0.683	0.667	4
SalGAN [36]	1.510	0.654	0.601	1.301	0.808	0.758	Baseline

Note: The rank of the team in the Saliency4ASD Grand Challenge is indicated in the last column. Specific details on the ranking method can be found in [13]. (↑: the higher the better, ↓: the lower the better).

TABLE 10.2 Performance for the set of classification metrics of the submitted models for classification of individuals between ASD and TD.

Team	Acc.	Recall	Precision	F1	Cohen's κ	AUC	Specificity	Rank
TUM [37]	0.598	0.717	0.574	0.632	0.201	0.644	0.484	1
R3U [39,40]	0.593	0.684	0.570	0.616	0.189	0.595	0.506	2
UM [43]	0.557	0.877	0.532	0.658	0.127	0.564	0.251	3
	0.579	0.592	0.563	0.570	0.158	0.579	0.566	
	0.574	0.594	0.568	0.568	0.149	0.575	0.556	
UCD&UK [44,45]	0.551	0.635	0.527	0.546	0.106	0.613	0.471	4
	0.542	0.741	0.522	0.610	0.091	0.575	0.351	
	0.539	0.807	0.519	0.629	0.089	0.544	0.282	
ECNU [47]	0.516	0.705	0.504	0.585	0.041	0.521	0.337	5
	0.446	0.397	0.429	0.412	−0.110	0.445	0.493	
	0.420	0.442	0.413	0.427	−0.159	0.421	0.399	
Random	0.499	0.499	0.488	0.489	−0.001	0.499	0.499	Baseline
ALL_ASD	0.489	1	0.489	0.656	0	0.5	0	Baseline
All_TD	0.511	0	0	0	0	0.5	1	Baseline

Note: The rank of the team in the Saliency4ASD Grand Challenge is indicated in the last column. Specific details on the ranking method can be found in [13]. (↑: the higher the better, ↓: the lower the better).

10.2.2 Face stimuli

10.2.2.1 Dataset

Duan et al. [19] further established a dataset of eye movements for children with autism on face stimuli, which is the first large-scale open source dataset of its kind. A total of 300 images selected from an open face dataset [50] were included in the dataset considering the balance of various emotions and whether the content is appropriate for children. The collected images contain faces

of various sizes, poses, emotions, ages, genders, etc. The selected images can be classified into six expressions (emotions), including generally positive, very positive, neutral, generally negative, very negative, and complex expressions, respectively, which are demonstrated in Fig. 10.4. Each expression has 50 images in this dataset. This dataset can be used to explore that the gaze patterns of children with autism and healthy children are similar or different for face stimuli under wild condition.

The procedure of recruiting subjects and conducting subjective experiments are similar as mentioned above. Please refer to [19] for more details.

10.2.2.2 Analysis

Face is a strong semantic region, which contains many salient features. To comprehensively compare the visual attention of individuals with autism and TD controls on face stimuli, the facial region was first divided into several regions of interest (ROI). The facial landmarks as well as pose and emotion were detected using a facial behavior analysis toolkit [51]. The CE-CLM approach [52] was adopted to localize 66 landmark points and estimate face pose, and the action unit (AU) detection system as mentioned in [53] was used to calculate facial expressions in this work. Then several regions including face region, ROI, and sub-ROI were defined accordingly which can be found in Fig. 10.5.

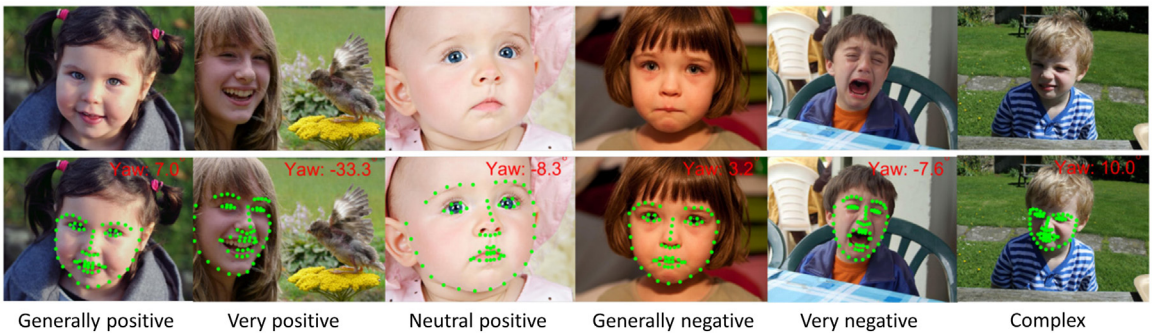


FIGURE 10.4 Face stimuli with various expressions. From the left to the right are images with generally positive, very positive, neutral, generally negative, very negative and complex expressions, respectively.

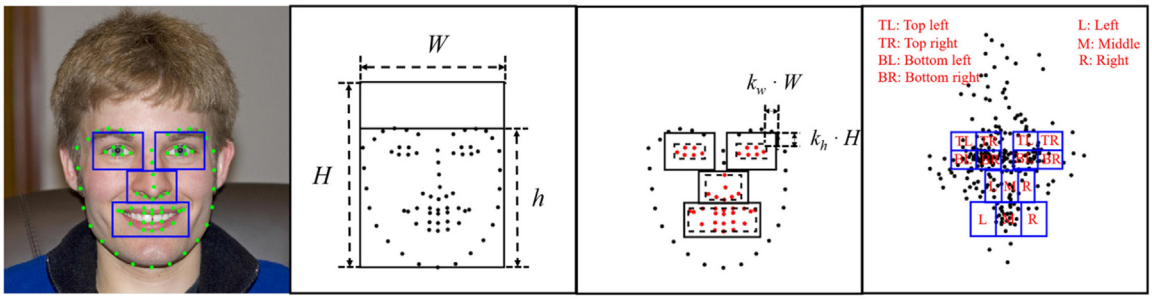


FIGURE 10.5 Definition of facial regions. (A) An example face image with facial landmarks. (B) The definition of facial region. (C) The definition of region of interest. (D) The definition of sub-region of interest.

Then, based on the collected eye movement data and label information obtained from facial behavior analysis toolkit, the analyses for the differences between autistic children and controls were conducted. First of all, it is necessary to analyze the visual attention difference on various regions. As shown in Fig. 10.6, the differences between the statistic information on each region are obvious. Fig. 10.6A illustrates the comparison of the percentage of fixation points on the facial region between autistic children and controls. The green points and blue points represent the percentage of fixation points on faces for TD controls and autistic children, respectively. The red points and black points are filtered data of green points and blue points, respectively. The filter is a moving average filter with the span of 30. A similar tendency can be observed with respect to the percentage of fixation points on human faces, which increases along with the increase of facial proportion in the image for both autistic children and controls. Meanwhile, TD controls focus more on human facial regions compared to autistic children.

Fig. 10.6B demonstrates the comparison of fixation proportion in each ROI for autistic children and TD controls. We can observe that the fixation proportion of children with autism in each ROI is less than that of controls. This may reveal the atypical visual attention of children with autism on human faces. Note that for the ASD group, the percentage of fixations in eyes ROIs is more than it is in nose or mouth ROIs. The hypothesis of excess mouth viewing in autism did not receive support in this study. This phenomenon is in-line with the majority of studies in the related field of ASD [18,54]. For nose and mouth regions, it can be observed that autistic children fixate almost same on two regions, while TD children focus more on nose regions compared to mouth regions.

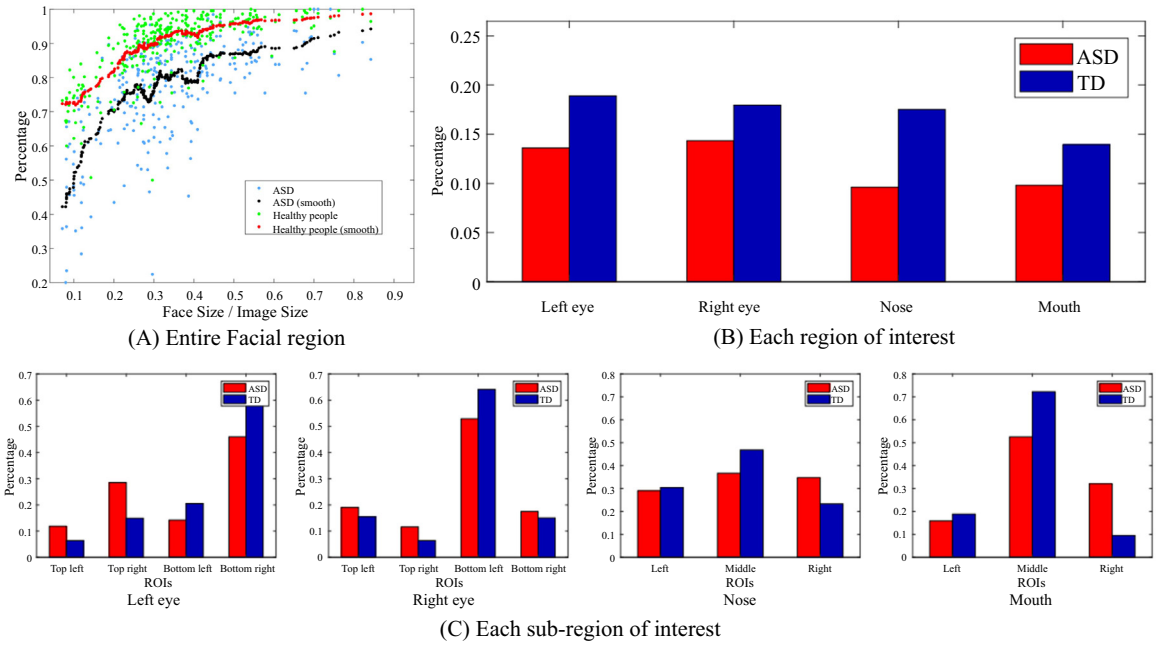


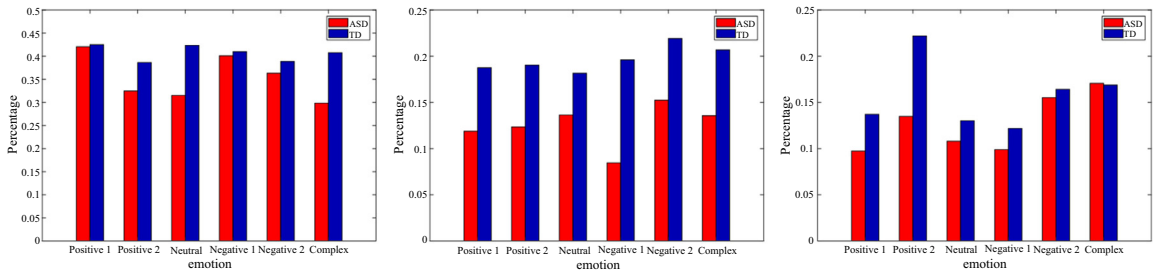
FIGURE 10.6 Comparison of fixation distribution between autistic children and healthy children on different regions. (A) On facial region. (B) In each region of interest. (C) In each sub-region of interest.

Fig. 10.6C illustrates the fixation distribution differences in each sub-ROI. As shown in Fig. 10.5, each eye region is segmented into four parts, including top left, top right, bottom left, bottom right, and each nose or mouth region is segmented into three parts, including left, middle, and right. Comparing the sub-regions in left eye and right eye, it can be observed that both autistic children and TD controls concentrate most at the bottom right and bottom left parts for left eye and right eye regions, respectively. For the second salient region, children with autism focus more at top right and top left regions for left eye and right eye regions, while TD children tend to look more at bottom left and bottom right regions for left eye and right eye regions. Comparing the sub-regions in nose and mouth regions, the middle parts of both nose and mouth regions are more salient than other parts. Moreover, the fixation distribution of children with autism is more dispersed compared to that of controls in the eye region, nose region or mouth region.

The influence of facial expressions on the visual attention of autistic children and TD controls were also analyzed. Fig. 10.7 illustrates the influence of facial expressions on the fixation distribution for autistic children and controls, respectively. "Positive 1" represents generally positive facial expressions (e.g., smile), "positive 2" represents very positive expressions (e.g., laugh), "neutral" represents neutral expressions, "negative 1" denotes generally negative facial expressions (e.g., sad), "negative 2" denotes very negative facial expressions (e.g., cry), "complex" represents complex expression (e.g., surprise). As shown in Fig. 10.6A, percentages of fixation in eye region are around 0.4 under all facial expressions for TD controls, while under "positive 2," "neutral," and complex facial expressions, children with autism focus less on eye region compared to other facial expressions. Fig. 10.6B illustrates the fixation percentages in the nose region under different facial expressions. It can be observed that TD controls have similar fixation percentages in the nose region under different facial expressions, while autistic children fixate less in the nose region under "negative 2" expression compared to other expressions. Fig. 10.6C demonstrates the fixation percentages in the mouth region under different facial expressions. It is obvious that both autistic children and TD children fixate less on the mouth region under "positive 1," "neutral," and "negative 1" expressions. For the rest expressions, controls focus more on the mouth region under "positive 2" expression, while autistic children fixate less under this expression.

10.2.2.3 *Methods and results*

Based on previously obtained atypical features of the visual attention of autistic children, a feature fusion network was designed to predict the gaze pattern of autism. To this end,



(A) Eye Region (B) Nose Region (C) Mouth Region

FIGURE 10.7 Influence of facial expressions on the fixation distribution in each region of interest. (A) In eye region. (B) In nose region. (C) In mouth region.

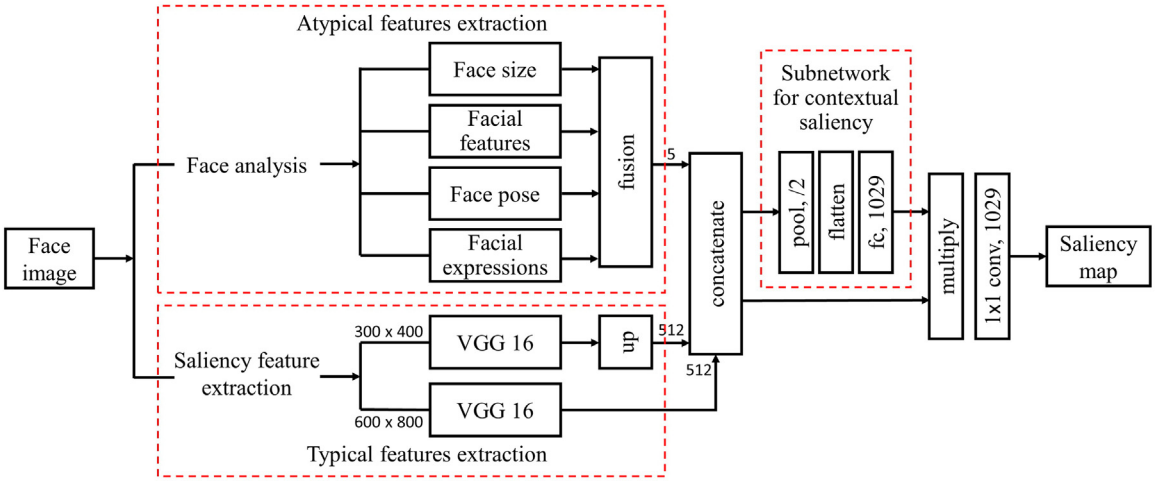


FIGURE 10.8 Atypical saliency prediction model of autism for face stimuli. The extracted atypical features are 5-dimension feature maps.

5-dimensional atypical features were extracted and fused and then concatenated to typical features extracted by two VGG-16 networks. Then with the help of a subnetwork for contextual saliency, the predicted saliency map can be obtained. The model is illustrated in Fig. 10.8.

As shown in Table 10.3, this specific model achieves the best performance among all competitive methods. Fig. 10.9 shows the comparison results of this method with the ground truths and the results of other methods. It can be observed that the differences between the visual attention map of ASD and TD are obvious. The visual attentions of autistic children are more dispersed. The fine-tuned CASNet [55] is better than the raw model. Moreover, with the combination of extracted atypical features, the model performs better on the prediction of the visual attention of autism on face stimuli.

TABLE 10.3 Comparison of different methods for predicting the visual attention of autistic children on face stimuli. All methods were fine-tuned on the dataset.

Model	AUC-Judd	sAUC	CC	NSS
SALICON	0.8087	0.5552	0.6448	1.4237
mlnet	0.8186	0.5598	0.6955	1.6011
SAM-VGG	0.8369	0.5644	0.7710	1.7594
SalGAN	0.8256	0.5752	0.7422	1.6811
CASNet [35]	0.8350	0.6166	0.7800	1.7492
Duan et al. [19]	0.8480	0.6232	0.8272	1.8239

10.2.3 Gaze-following stimuli

10.2.3.1 Dataset

As mentioned in the introduction, individuals with autism show atypical joint attention behavior, which motivated a study of exploring the visual attention of them on gaze-following stimuli [23]. A GazeFollow4ASD dataset was established which includes 300 images with gaze-following cues in the wild. These images were collected from GazeFollow dataset [56], which is a large-scale dataset with annotations of the location of eyes of people and where they are looking at (i.e., gazed objects).

Eight high-functioning autistic children were recruited, whose ages were between 4 and 16 years with an average of 9.6 years. Their IQs were assessed using the Wechsler

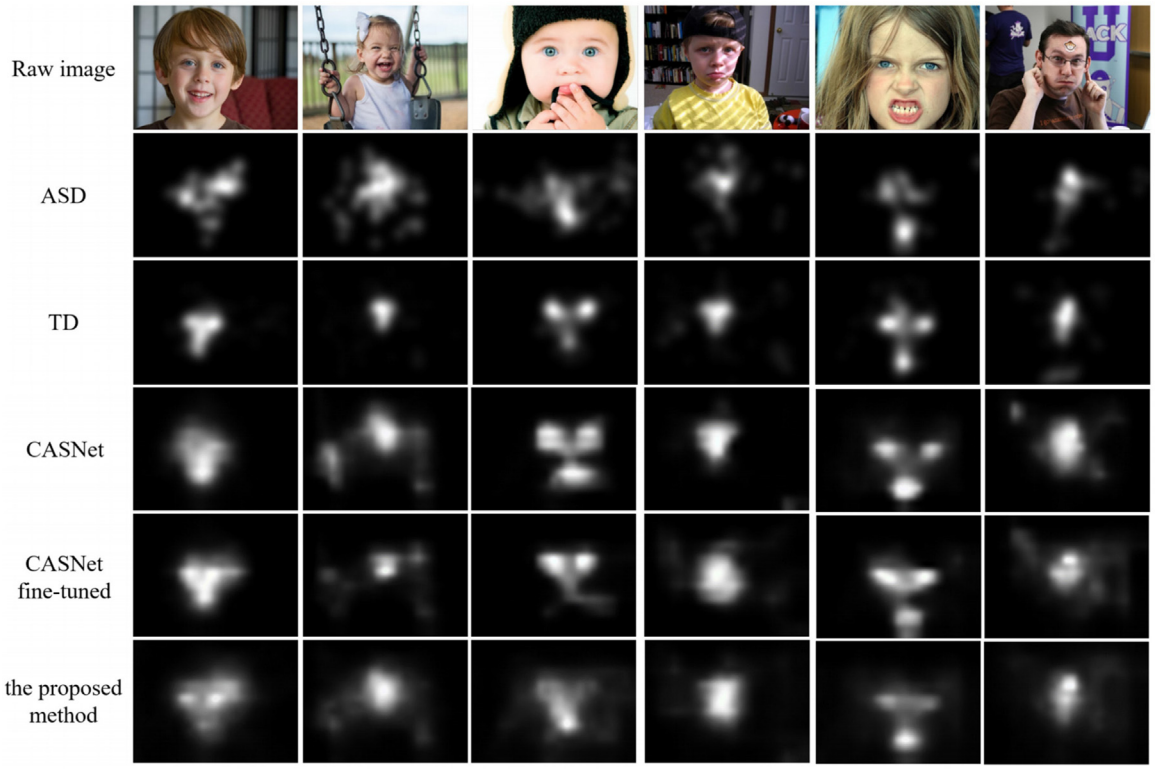


FIGURE 10.9 Visual attention map and predicted saliency map. From the first row to the bottom row are raw images, visual attention map for autistic children, visual attention map for TD children, saliency map predicted by CASNet, saliency map predicted by fine-tuned CASNet, saliency map predicted by the proposed model, respectively.

Abbreviated Scale of Intelligence (WASI) before the experiment, which were greater than 70. All children with autism met the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Ten healthy controls without psychiatric history or developmental delay were recruited correspondingly, whose ages were between 4 and 14 years with an average of 8.9, and WISC-IV IQs were greater than 70. The gender and education were also matched between two groups. Before the experiment, all subjects were confirmed to have normal or corrected-to-normal visual acuity. The procedure of the experiment was similar with that mentioned above, reader can also refer to [23] for more details.

10.2.3.2 Analysis

Fig. 10.10 shows the comparisons between the visual attention maps of autistic children and TD controls on gaze-following stimuli. As shown in Fig. 10.10A, when there is only one gaze interaction in an image, which is the simplest scenario, TD controls tend to pay more attention on both eyes and gaze-at objects and focus more on eyes or faces, while autistic children prefer to focus the gaze-at objects. As shown in Fig. 10.10B, when the situation is more complicated, that is, several gaze interactions in an image, TD controls prefer to focus eyes and faces of the main character in the image for social reference, while autistic children tend to look more at the gaze-at objects.

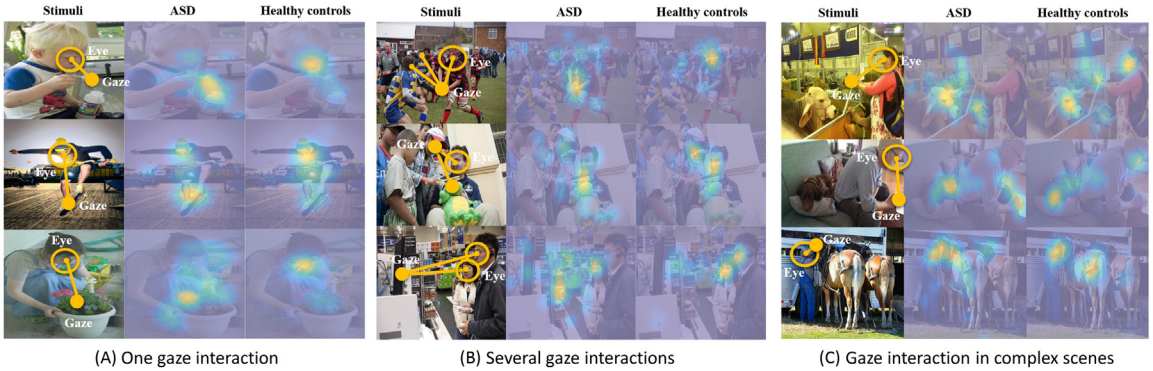


FIGURE 10.10 The Comparisons between the visual attention maps of autistic children and TD controls on gaze-following stimuli. (A) One gaze interaction in an image. (B) Several gaze interactions in an image. (C) Gaze interaction in complex scenes. Three columns from the left to the right represent the image stimuli, the visual attention map for children with autism and the visual attention map for TD controls, respectively.

TABLE 10.4 The comparison of two datasets (Jiang et al. and Fang et al.) considering the similarity (i.e., difference) between the visual attention maps for children with autism and healthy controls.

Dataset	Metric AUC-Judd	sAUC	CC	NSS
Jiang et al. [57]	0.9185	0.8296	0.9571	2.6538
Fang et al. [23]	0.8438	0.6692	0.7466	2.0547

This special gaze-following phenomenon further proves the lack of social interactions monitoring in ASD. As shown in Fig. 10.10C, when the gaze-following behavior occurs in a more complex scene or is not the main part of the image, children with autism show pixel-level or object-level visual attention, while TD controls show more semantic attention.

A comparison study between the gaze-following stimuli and natural stimuli was conducted. Table 10.4 shows the comparison between two datasets (Jiang et al. [57] and Fang et al. [23]), which conducted eye movement studies based on natural image stimuli and gaze-following stimuli, respectively. The saliency metrics AUC, sAUC, CC, and NSS were used to compare the similarity (i.e., difference) between the visual attention map of individuals with autism and controls. It can be observed that the difference between two groups in the dataset of Jiang et al. [57] is less than that in the dataset of Fang et al. [23], which may indicate that the gaze-following stimuli are more discriminative for distinguishing two groups (i.e., individuals with autism and controls) based on eye movements.

10.2.3.3 *Methods and results*

Two subtasks were needed to classify the gaze patterns between individuals with autism and controls. The first subtask was to extract discriminative features based on the visual attention map of autism group and control group. To highlight the difference between the visual attention of autism group and control group, the differences of fixation (DoF) density maps [57] were obtained by calculating the normalized pixel-wise subtraction of fixation maps for two groups as:

$$DoF = F^{+} - F^{-}, \tag{10.1}$$

where F^{+} and F^{-} denote the fixation density maps for individuals with autism and controls, respectively. In this way, the visual attention maps of two groups can be combined together,

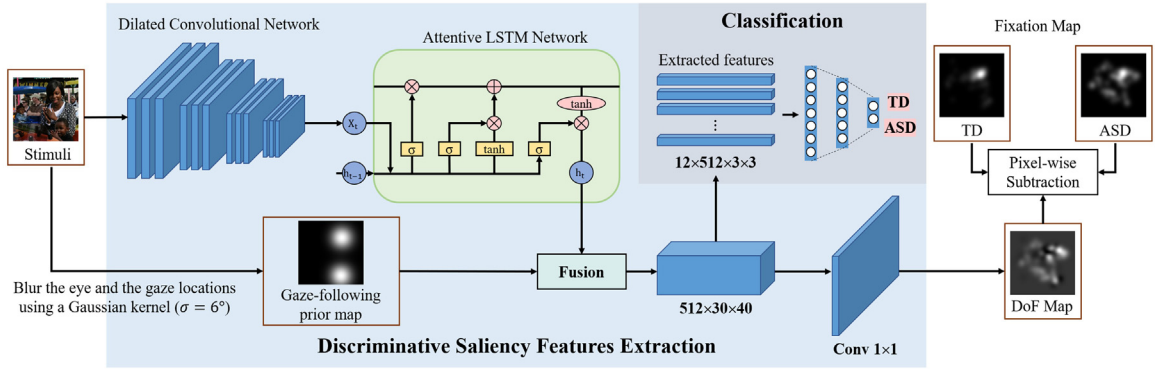


FIGURE 10.11 A LSTM-based model for extracting atypical features of autism by predicting the difference of fixation (DoF) maps. Then this model is used to classify the gaze patterns of individuals with autism and healthy people.

which enables that distinguishing zones are highlighted and similar zones are masked.

Then a discriminative saliency-features extraction module was designed to extract atypical DoF features as shown in Fig. 10.11. A dilated convolutional network was adopted to extract coarse feature maps from the input images, which was modified from ResNet-50 by removing the stride and introducing dilated convolutions in the last two blocks. A long short-term memory (LSTM) module was followed to recurrently process the extracted features and enhance the saliency prediction. Finally, the gaze-following prior was adopted as the visual attention bias, instead of the center bias, in the fusion step to further highlight the gaze-following semantic information.

Finally, the Pearson's Correlation Coefficient (CC) and Kullback–Leibler divergence (KL) functions were adopted as the loss function for predicting the saliency DoF map, which is defined as:

$$L(P,G) = \frac{\sigma(P,G)}{\sigma(P)\times\sigma(G)} + \sum_i G_i \log\left(\varepsilon + \frac{G_i}{\varepsilon + P_i}\right), \tag{10.2}$$

where P and G represent the predicted DoF map and ground-truth DoF map, respectively, σ is the covariance, i denotes the i-th pixel, and ε is the regularization constant.

TABLE 10.5 Performance comparison between the classification accuracy of two methods (Jiang et al. and Fang et al.) on two datasets.

Dataset	Method		
	Jiang et al.	Zero padded features (%)	Fang et al.
Jiang et al. [57]	52.01%	70.34	71.63%
Fang et al. [23]	56.82%	77.12	79.94%

For the second subtask, that is, classification task, a specific classification module was designed. First, the feature vector was extracted from the feature map extracted as mentioned above at each location of fixation point. Only the feature vectors corresponding to the first 12 fixation points were extracted with the dimension of 512a1 × 1 for each feature vector. After passing through two fully connected (FC) layers with units of 1024 and 128 respectively and a softmax classifier, the final prediction can be obtained.

As shown in Table 10.5, this method achieved the best performance compared to other methods on both the natural stimuli and the gaze-following stimuli. Moreover, comparing the performance of three methods between two datasets, it can be observed that all models can achieve better classification results on this

gaze-following dataset, which indicates that the gaze-following stimuli (i.e., joint attention stimuli) may achieve better discriminative ability between the gaze patterns of individuals with autism and healthy people.

10.3 Action behavior phenotype

Besides eye movements, individuals with autism also show other types of atypical behavior. Although these behaviors may be different or even nonexistent among individuals with autism, they may help to judge the degree of autism or assist the diagnosis from another perspective. In this section, we discuss the atypical behavior features of autism and automatic models for recognizing them.

10.3.1 Dataset and analysis

A total of 30 videos (about 40 h) compose the ASD40h dataset, which contains 5 most common autistic behaviors, including hand flapping, head banging, spinning in a circle, toe walking, and moving fingers. Fig. 10.12 shows the example of these 5 common atypical behaviors. Atypical action instances and repetitive behavior instances were annotated in each video.

The dataset was divided into 20 training and 10 testing sequences. Data augmentation manipulations were performed before training, which includes segmenting the person and scene background to increase the diversity of the background and using shear transformation to efficiently simulate limited viewpoint changes. Random flipping and corner cropping like manipulations were also performed to further augment the video data.

10.3.2 Methods and results

Fig. 10.13 presents the overview of the system. To perform atypical action behavior recognition, a specific model was designed, which includes 4 components as shown in Fig. 10.14, that is, a 3D ConvNet temporal feature extractor, a temporal pyramid network, an *ASD action detector*, and a repetitive behavior discriminator. To enable efficient computation and end-to-end training, the C3D temporal feature maps were shared in two tasks. The temporal pyramid network was used to shorten the feature map, while keeping the high-level semantics. The *ASD action detector* was used to

FIGURE 10.12 Atypical action behavior for children with autism, including hand flapping, head banging, spinning in a circle, toe walking, and moving fingers, etc.

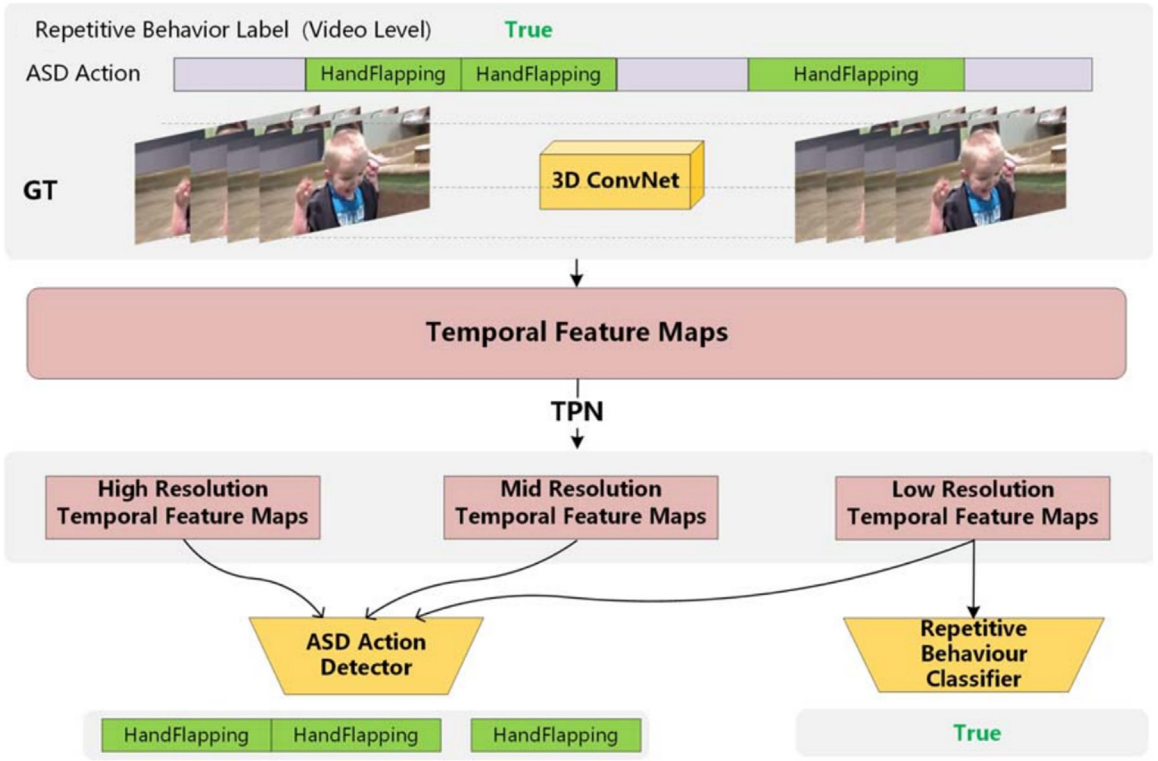


FIGURE 10.13 Overview of the system.

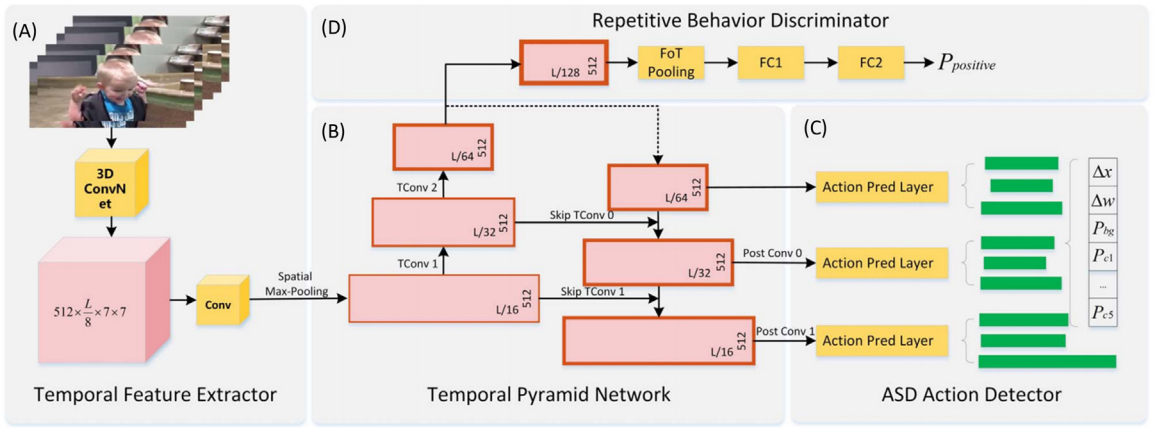


FIGURE 10.14 Framework of the approach. (A) Temporal feature extractor module. (B) Temporal pyramid network. (C) Action detector for autism. (D) Repetitive behavior discriminator.

classify the video segments into multiple categories. The repetitive behavior discriminator was used to discriminate whether the whole video contains repetitive behavior.

The first part of the model is the video—temporal feature extractor. A 3D ConvNet was used to extract spatio-temporal feature from the assigned video buffer, of which the structure is similar to the C3D architecture. To obtain temporal-only features, a 3D convolutional filter was added to extend the temporal receptive field. A 3D max-pooling filter was adopted to downsample the spatial dimensions of the features.

A fully-convolutional temporal pyramid network (TPN) similar to feature pyramid network (FPN) was adopted to utilize high-level semantic temporal features at all scales. It includes a top-down pathway and a bottom-up pathway with skip connections to build relationship. The bottom-up pathway was feed-forward computation of anchor layers. The top-down pathway generated higher resolution features by upsampling temporally features, then, these features were enhanced from the bottom-up pathway via skip connections. Three anchor features were finally obtained.

After extracting features, an *ASD action detector* module was followed to detect the atypical action behavior. For each temporal feature map of anchor layers, specific actions of autism were assigned. For the lower anchor layers, which had smaller resolution and larger receptive field, it was used to predict long ASD action (e.g., moving fingers in front of the eyes). For the top anchor layers, which had larger resolution and smaller receptive field, it was used to predict short ASD action (e.g., hand flapping).

Finally, a *repetitive behavior discriminator* module was designed to classify whether the video contains individuals with repetitive behavior. This discrimination of repetitive behaviors can be seen as a classification task. The high-level semantic features were used for this task. A fixed-length temporal pooling layer, named FoT pooling, was designed to extract the fixed-length temporal features from the top high-level semantic features. The output of the FoT pooling was then fed into two FC layers and then classified to predict whether this video contains repetitive behavior or not.

The training objective of this network was to solve a mulit-task optimization problem. The overall loss function was defined as:

$$L = L_{\text{action class}} + \alpha L_{action\ loc} + \beta L_{\text{repetitive class}}, \tag{10.3}$$

where $L_{\text{action class}}$ is the action classification loss for autism, $L_{action\ loc}$ is the action detection loss function for autism, $L_{\text{repetitive class}}$ is the probability loss function for the repetitive behavior, α and β are two balanced hyper parameters. Readers can refer to [26] for more details of these loss functions.

Table 10.6 shows the results of mAP comparisons of atypical action recognition task. It can be observed that the method of Tian et al. [26] achieved the best performance compared

TABLE 10.6 mAP results for action detection.

Method		S-CNN [58]	SST [59]	SSAD [60]	R-C3D [61]	Tian et al. [26]
α	0.1	49.3	50.4	53.0	58.3	56.6
	0.2	45.3	47.4	50.4	53.6	54.2
	0.3	37.8	44.3	47.8	48.2	50.1
	0.4	30.1	32.7	34.5	34.8	35.2
	0.5	21.8	25.5	25.7	27.2	28.1

to other competitors surpassing the best state-of-the-art model by almost 1%. As for the repetitive behavior discrimination task, compared to 4 state-of-the-art methods widely used in this task, this method also achieved the best performance with the repetitive behavior classification accuracy up to 95.2% (Table 10.7).

However, though all these action recognition model can achieve good classification or prediction performance, whether it can be used in assisting diagnosis for autism is still need to be discussed. There are two reasons that may cause the false diagnosis. The first reason is that not all individuals with autism show atypical action behavior or they may show different atypical action behavior besides the dataset in [26], which may cause missed diagnosis. The second reason is that other diseases may also show atypical actions such as repetitive behavior and this may cause false diagnosis. However, this new method still has its possibility for screening ASD due to its large-range feasibility.

10.4 Drawing behavior phenotype

Besides atypical action behaviors, another behavior which is less studied (i.e., drawing behavior), may also reveal the hallmarks of autism. In this section, we will discuss this drawing behavior phenotype and its possible applications.

10.4.1 Dataset

Shi et al. [28] established a painting dataset drawing by individuals with ASD, which includes 478 paintings from 15 children with autism and 490 paintings from 20 TD children. These paintings can be classified into four categories, including barbola painting, line-drawing, oil painting, and watercolor painting, as illustrated in Fig. 10.15. Then a subjective experiment was conducted to extract manual features. Seven features inspired from the observation of painting differences between autistic children and TD controls, as well as the hallmarks of autism, were proposed, and then three clinicians were recruited to label these features. All three clinicians had never accessed to these paintings before the experiment.

10.4.2 Analysis

Two common atypical behaviors including atypical face processing behaviors and repetitive behaviors were first analyzed. As reported in [19], individuals with autism have impairments in face recognition or discrimination

TABLE 10.7 Performance of repetitive behavior recognition task.

Method	iDT [62]	Two-stream [63]	C3D (v1) [64]	C3D (v2) [64]	Tian et al. [26]
Accuracy (%)	88.2	89.0	93.6	94.5	95.2

barbola Line-drawing oil watercolor

FIGURE 10.15 Examples of four categories of paintings in the PASD dataset.

compared to TD individuals, which may denote the atypical facial semantic processing of autism. As shown in the left part of Fig. 10.16, comparing the face paintings from children with autism and TD children, the face paintings from children with autism are more abstract and weird. Moreover, quantitative comparison between the number of face paintings demonstrates children with autism may draw more paintings with faces. When comparing the average face numbers in their paintings, it can be observed that children with autism often draw more faces in one painting compared to TD children. Furthermore, repetitive behaviors can also be observed in the paintings of autism. As shown in the right part of Fig. 10.16, children with autism draw more paintings with abstract repetitive patterns. The quantitative comparison indicates the significance of this difference. As for the number of the repetitive patterns in one painting, there is no difference between the two groups.

In addition to the two aforementioned atypical behaviors, four specific features for paintings were also analyzed, as shown in Fig. 10.17. Comparing the logical structure

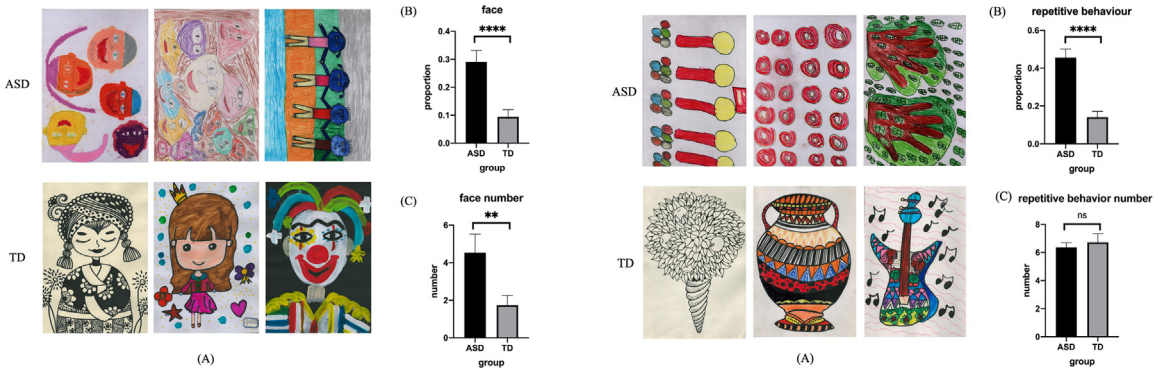


FIGURE 10.16 Comparisons of two widely studied behaviors, that is, face processing behaviors and repetitive behaviors, between paintings of autistic children and healthy controls.

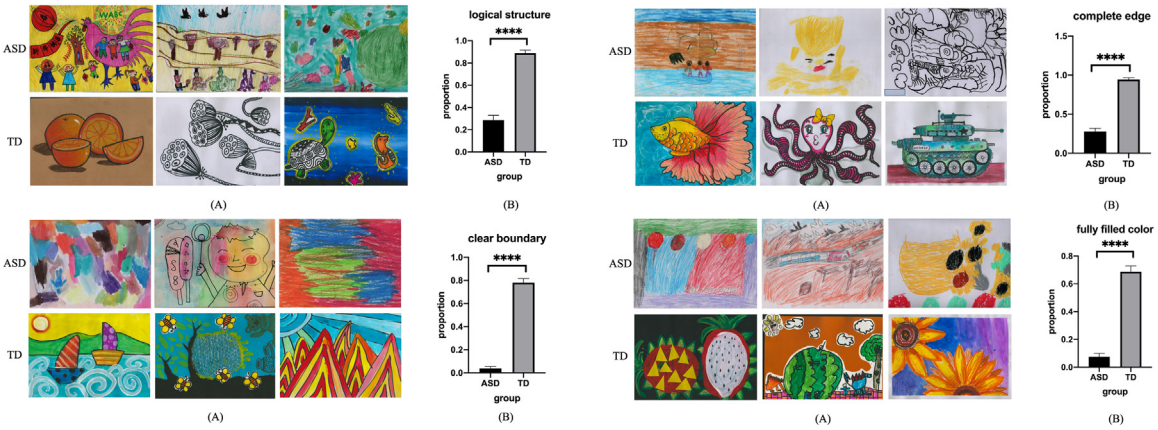


FIGURE 10.17 Comparisons of four specific painting behaviors, including logical structure, complete edge, clear boundary, as well as fully filled color, between paintings of autistic children and healthy controls.

FIGURE 10.18 Comparisons of composition location between paintings of autistic children and healthy controls.

between the paintings of autistic children and healthy children, it can be observed that paintings from children with autism barely follow the normal logic, for example, people are drawn upon a rooster or other meaningless combinations. Comparing the edges in paintings, children with autism usually draw paintings without complete edges while TD children often draw clear and complete borderline. As for the boundary between different parts or objects in paintings, children with autism may not draw paintings with clear boundary while TD controls tend to draw paintings with clear demarcation of colors or objects. Moreover, paintings from children with autism may have large blank gap without fully filled color. Finally, as shown in Fig. 10.18, children with autism usually do not consider the composition structure of the painting while TD children tend to draw their paintings symmetrically and centered.

10.4.3 Results and discussion

Based on above analysis and extracted features, a SVM classification experiment was conducted. Table 10.8 shows the results of the classification results.

TABLE 10.8 Classification results of SVM model on this PASD dataset.

	Median	Average	Variance
Accuracy	0.8981	0.8519	0.0089

From the above analysis and results, it can be observed that the difference between the paintings from children with autism and TD people are obvious. Although whether it can be used in the procedure of aiding the diagnosis still needs further discussion, since it is easy to obtain paintings from children for parents, this method may be helpful for the large-scale screening of autism.

10.5 Discussion and conclusion

In this paper, we have discussed several behavioral phenotype features of autism and several state-of-the-art techniques related to automatically and quantitatively measuring these phenotypic features. The phenotype markers of autism mainly include social communication symptoms, fixated or restricted behaviors or interests, hyper- or hypo-sensitivity to sensory stimuli,

and associated features [3]. Specifically, in this chapter we mainly discussed three types of behavior features of individuals with autism including eye movement behavior, action behavior, as well as drawing behavior, which can reflect the phenotype symptoms of atypical social communication and fixated or restricted behaviors of autism. In this section, we will briefly summarize this chapter and discuss several issues that need to be explored in the future, in order to better use computer-related technologies to promote the progress of autism community.

As mentioned above, eye movements of individuals with autism are significantly different to typically developed individuals. We have compared the difference of eye movement patterns on three types of stimuli including stimuli in the wild, face stimuli, and gaze-following stimuli. Significant differences can be observed in all these three types of stimuli, including social or nonsocial stimuli. However, as discussed in [23], social stimuli, such as gaze-following, may better help to classify the gaze patterns between autism and typical development. The natural stimuli in the wild may be better for defining special traits for autism. As for facial stimuli, facial expressions have significant impacts on the visual attention of autism [19]. However, single facial in one image is not strong enough to differentiate the state of autism. We argue that incorporating the facial stimuli with various expressions into strong social stimuli (such as gaze-following) may better help to distinguish the gaze patterns between individuals with autism and TD people. Moreover, since videos have more social stimulation, they may better help distinguish two groups. However, few studies have been conducted on such stimuli. And since the videos can incorporate audio stimuli into it to guide the visual attention, mores studies are still needed to analyze whether it is significant to use and how to design audio stimuli. Another issue that has not been considered in previous studies is the viewport of stimuli. Previous studies use the 2D screen to display stimuli and attract visual attention, which may cause two problems. First of all, individuals may not concentrate on the screen and will move their eyes off the screen, which may cause the failure of eye tracking and add additional data collection workload. This may influence the application of it into diagnosis procedure. Furthermore, restricted viewport is different with the real-world perception, which may not comprehensively reflect the visual traits of autism. Future studies using Virtual Reality (VR) or CAVE [65] contents as stimuli may help quantify the traits and aid the diagnosis of autism better.

Atypical action behaviors are also the core symptoms of autism, which attributes to the fixated or restricted behavior phenotype characteristics. The model in reference [26] can accurately classify the actions of individuals with autism and TD individuals. However, atypical actions of individuals with autism are complex and diverse for different individuals, which requires more research and discussion on this issue. As for the repetitive action behavior, since not only children with autism, but also individuals with other disorders may also exhibit this atypical behavior, how to use it still needs more future studies. Moreover, drawing behaviors can also reveal several hallmarks of autism, since painting can not only show action behaviors, but also reflect the human cognition, and psychological factors. However, the work in [28] needs to manually label several factors, automatic classification of these paintings requires further research. These action-based methods do not need special stimuli and can be easily obtained, which makes these methods can be used to screen children with autism.

Overall, using state-of-the-art AI techniques (e.g., computer vision, natural language processing, machine learning, etc.) with appropriate stimuli (e.g., social stimuli, video stimuli, etc.) to assist the screening and diagnosis of

autism is significant and possible, and more exploration studies should be conducted. We hope this paper can help other researchers conduct experiments and facilitate future studies related to this topic.

References

[1] S. Baron-Cohen, A.M. Leslie, U. Frith, Does the autistic child have a "theory of mind"? Cognition 21 (1985) 37–46.

[2] S. Baron-Cohen, S. Wheelwright, The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences, Journal of Autism and Developmental Disorders 34 (2004) 163–175.

[3] F. Edition, Diagnostic and statistical manual of mental disorders, American Psychiatric Association (2013) 21.

[4] C.E. Robertson, S. Baron-Cohen, Sensory perception in autism, Nature Reviews. Neuroscience 18 (2017) 671–684. Available from: https://doi.org/10.1038/nrn.2017.112.

[5] S.D. Tomchek, W. Dunn, Sensory processing in children with and without autism: a comparative study using the short sensory profile, American Journal of occupational therapy 61 (2007) 190–200.

[6] D.R. Simmons, A.E. Robertson, L.S. McKay, E. Toal, P. McAleer, F.E. Pollick, Vision in autism spectrum disorders, Vision Research 49 (2009) 2705–2739.

[7] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, et al., A dataset of eye movements for the children with autism spectrum disorder. In Proceedings of the ACM Multimedia Systems Conference, 2019:255–260.

[8] J. Osterling, G. Dawson, Early recognition of children with autism: A study of first birthday home videotapes, Journal of Autism and Developmental Disorders 24 (3) (1994) 247–257.

[9] K. Chawarska, S. Macari, F. Shic, Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders, Biological Psychiatry 74 (3) (2013) 195–203.

[10] S. Wang, M. Jiang, X.M. Duchesne, E.A. Laugeson, D.P. Kennedy, R. Adolphs, et al., Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking, Neuron 88 (3) (2015) 604–616.

[11] H. Duan, G. Zhai, X. Min, Y. Fang, Z. Che, X. Yang, et al., Learning to predict where the children with asd look. In Proceedings of the IEEE International Conference on Image Processing; 2018:704–708.

[12] C. Wu, S. Liaqat, H. Duan, S. Ozonoff, C.N. Chuah, G. Zhai, et al., Machine-Learning Based Autism Diagnosis Using Gaze Fixations on Natural Images. In Proceedings of the INSAR 2020 Virtual Meeting.

[13] J. Gutiérrez, Z. Che, G. Zhai, P. Le Callet, Saliency4ASD: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder, Signal Processing: Image Communication 92 (2021) 116092.

[14] C. Katarzyna, V. Fred, Limited attentional bias for faces in toddlers with autism spectrum disorders, Archives of General Psychiatry 67 (2) (2010) 178–185.

[15] T. Falck-Ytter, C. von Hofsten, How special is social looking in ASD: a review, Progress in Brain Research 189 (2011) 209–222.

[16] B. Corden, R. Chilvers, D. Skuse, Avoidance of emotionally arousing stimuli predicts social–perceptual impairment in Asperger's syndrome, Neuropsychologia 46 (1) (2008) 137–147.

[17] Y. Bar-Haim, C. Shulman, D. Lamy, A. Reuveni, Attention to eyes and mouth in high-functioning children with autism, Journal of Autism and Developmental Disorders 36 (1) (2006) 131–137.

[18] J. Åsberg Johnels, D. Hovey, N. Zürcher, L. Hippolyte, E. Lemonnier, C. Gillberg, et al., Autism and emotional face-viewing, Autism Research 10 (5) (2017) 901–910.

[19] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, G. Zhai, Visual attention analysis and prediction on human faces for children with autism spectrum disorder, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15 (3s) (2019) 1–23.

[20] S.C. Butler, A.J. Caron, R. Brooks, Infant understanding of the referential nature of looking, Journal of COgnition and Development 1 (4) (2000) 359–377.

[21] T. Falck-Ytter, E. Fernell, Å.L. Hedvall, C. Von Hofsten, C. Gillberg, Gaze performance in children with autism spectrum disorder when observing communicative actions, Journal of Autism and Developmental Disorders 42 (10) (2012) 2236–2245.

[22] M.R. Swanson, M. Siller, Patterns of gaze behavior during an eye-tracking measure of joint attention in typically developing children and children with autism spectrum disorder, Research in Autism Spectrum Disorders 7 (9) (2013) 1087–1096.

[23] Y. Fang, H. Duan, F. Shi, X. Min, G. Zhai, Identifying Children with Autism Spectrum Disorder Based on Gaze-Following. In Proceedings of the IEEE International Conference on Image Processing; 2020:423–427.

[24] A. Klin, D.J. Lin, P. Gorrindo, G. Ramsay, W. Jones, Two-year-olds with autism orient to non-social contingencies rather than biological motion, Nature 459 (7244) (2009) 257–261.

[25] L. Fan, W. Cao, H. Duan, Y. Du, J. Chen, S. Hou, et al. Screening of Autism Spectrum Disorder Using Novel Biological Motion Stimuli. In Proceedings of the International Forum of Digital TV and Wireless Multimedia Communication, 2020:371.

[26] Y. Tian, X. Min, G. Zhai, Z. Gao, Video-based early asd detection via temporal pyramid networks. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2019:272–277.
[27] M.J. Emery, Art therapy as an intervention for autism, Art Therapy 21 (3) (2004) 143–147.
[28] F. Shi, W. Sun, H. Duan, X. Liu, M. Hu, W. Wang, et al., Drawing Reveals Hallmarks of Children with Autism, Displays (2021) 102000.
[29] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look. In Proceedings of the IEEE International Conference on computer vision, 2009:2106–2113.
[30] W. Wei, Z. Liu, L. Huang, A. Nebout, O. Le Meur, Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019: 621–624.
[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, arXiv:1409.1556.
[32] A. Nebout, W. Wei, Z. Liu, L. Huang, O. Le Meur, Predicting saliency maps for ASD people. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019:629–632.
[33] Y. Fang, H. Huang, B. Wan, Y. Zuo, Visual attention modeling for autism spectrum disorder by semantic features. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019:625–628.
[34] Y. Fang, H. Zhang, Y. Zuo, W. Jiang, H. Huang, J. Yan, Visual attention prediction for autism spectrum disorder with hierarchical semantic fusion, Signal Processing: Image Communication 92 (Apr. 2021) 116186.
[35] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional Networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015:234–241.
[36] J. Pan, C. Canton, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, et al. SalGAN: VIsual saliency prediction with generative adversarial networks, 2017, arXiv:1701.01081.
[37] M. Startsev, M. Dorr, Classifying autism spectrum disorder based on scanpaths and saliency. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019:633–636.
[38] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an LSTM-based saliency attentive model, IEEE Transactions on Image Processing 27 (10) (2018) 142–5154.
[39] G. Arru, P. Mazumdar, F. Battisti, Exploiting visual behaviour for autism spectrum disorder identification. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019:637–640.
[40] P. Mazumdar, G. Arru, F. Battisti, Early detection of children with autism spectrum disorder based on visual exploration of images, Signal Processing: Image Communication 92 (Apr. 2021) 116184.
[41] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:7263–7271.
[42] L. Zhang, Z. Gu, H. Li, SDSP: A novel saliency detection method by combining simple priors. In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2013:171–175.
[43] Y. Tao, M.-L. Shyu, SP-ASDNet: CNN-LSTM Based ASD classification model using observer scanpaths. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019:641–646.
[44] C. Wu, S. Liaqat, S.-C,S. Cheung, C.-N. Chuah, S. Ozonoff, Predicting autism diagnosis using image with fixations and synthetic saccade patterns. In Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019:647–650.
[45] S. Liaqat, C. Wu, P.R. Duggirala, S.-C.S. Cheung, C.-N. Chuah, S. Ozonoff, et al., Predicting ASD diagnosis in children with synthetic and image-based eye gaze data, Signal Processing: Image Communication 92 (2021) 116198.
[46] C. Wloka, I. Kotseruba, J.K. Tsotsos, Saccade sequence prediction: Beyond static saliency maps, 2017, arXiv preprint arXiv:1711.10959.
[47] S. Xu, J. Yan, M. Hu, A new bio-inspired metric based on eye movement data for classifying ASD and typically developing children, Signal Processing: Image Communication 92 (2021) 116171.
[48] A. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (3) (2019) 740–757.
[49] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (4) (2009) 427–437.
[50] X. Min, G. Zhai, K. Gu, J. Liu, S. Wang, X. Zhang, et al., Visual attention analysis and prediction on human faces, Information Sciences 420 (2017) 417–430.
[51] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.P. Morency, Openface 2.0: Facial behavior analysis toolkit. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018:59–66.

[52] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, L.P. Morency, Convolutional experts constrained local model for 3D facial landmark detection. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017:2519–2528.
[53] T. Baltrušaitis, M. Mahmoud, P. Robinson, Cross-dataset learning and person-specific normalisation for automatic action unit detection. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015 6:1–6.
[54] T. Falck-Ytter, S. Bölte, G. Gredebäck, Eye tracking in early autism research, Journal of Neurodevelopmental Disorders 5 (1) (2013) 1–13.
[55] S. Fan, Z. Shen, M. Jiang, B.L. Koenig, J. Xu, M.S. Kankanhalli, et al., Emotional attention: A study of image sentiment and visual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:7521–7531.
[56] A. Recasens, A. Khosla, C. Vondrick, A. Torralba, Where are they looking? In NeurIPS 2015.
[57] M. Jiang, Q. Zhao, Learning visual attention to identify people with autism spectrum disorder. In Proceedings of the IEEE International Conference on Computer Vision, 2017:3267–3276.
[58] Z. Shou, D. Wang, S.F. Chang, Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016:1049–1058.
[59] S. Buch, V. Escorcia, C. Shen, B. Ghanem, J. Carlos Niebles, Sst: Single-stream temporal action proposals. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017:2911–2920.
[60] T. Lin, X. Zhao, Z. Shou, Single shot temporal action detection. In Proceedings of the ACM International Conference on Multimedia, 2017:988–996.
[61] H. Xu, A. Das, K. Saenko, R-c3d: Region convolutional 3d network for temporal activity detection. In Proceedings of the IEEE International Conference on Computer Vision, 2017:5783–5792.
[62] H. Wang, C. Schmid, Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, 2013:3551–3558.
[63] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos. In NeurIPS 2014.
[64] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, 2015:4489–4497.
[65] M.E. Minissi, I.A.C. Giglioli, F. Mantovani, M.A. Raya, Assessment of the autism spectrum disorder based on machine learning and social visual attention: a systematic review, Journal of Autism and Developmental Disorders (2021) 1–16.