Gaze Patterns in Children With Autism Spectrum Disorder to Emotional Faces: Scanpath and Similarity

Wei Zhou, Minqiang Yang, *Member, IEEE*, Jingsheng Tang, Juan Wang, and Bin Hu, *Fellow, IEEE*

***Abstract*— Autism spectrum disorder (ASD) one of the fastest-growing diseases in the world is a group of neurodevelopmental disorders. Eye movement as a biomarker and clinical manifestation represents unconscious brain processes that can objectively disclose abnormal eye fixation of ASD. With the aid of eye-tracking technology, plentiful methods that identify ASD based on eye movements have been developed, but there are rarely works specifically for scanpaths. Scanpaths as visual representations describe eye movement dynamics on stimuli. In this paper, we propose a scanpath-based ASD detection method, which aims to learn the atypical visual pattern of ASD through continuous dynamic changes in gaze distribution. We extract four sequence features from scanpaths that represent changes and the differences in feature space and gaze behavior patterns between ASD and typical development (TD) are explored based on two similarity measures, multimatch and dynamic time warping (DTW). It indicates that ASD children show more individual specificity, while normal children tend to develop similar visual patterns. The most noticeable contrasts lie in the duration of attention and the spatial distribution of visual attention along the vertical direction. Classification is performed using Long Short-Term Memory (LSTM) network with different structures and variants. The experimental results show that LSTM network outperforms traditional machine learning methods.**

***Index Terms*— Autism spectrum disorder, eye tracking, LSTM network, scanpaths.**

Manuscript received 7 May 2023; revised 1 December 2023 and 22 January 2024; accepted 26 January 2024. Date of publication 5 February 2024; date of current version 22 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200; in part by the National Natural Science Foundation of China under Grant 62227807; in part by the Natural Science Foundation of Gansu Province, China, under Grant 22JR5RA488; in part by the Fundamental Research Funds for the Central Universities under Grant lzujbky-2023-16; in part by the Key Research and Development Plan of Jining under Grant 2021YXNS040; and in part by the Supercomputing Center of Lanzhou University. *(Corresponding authors: Minqiang Yang; Bin Hu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Sichuan Guangyuan Mental Health Center under Application No. GJWLP20210011.

Wei Zhou, Minqiang Yang, Jingsheng Tang, and Bin Hu are with the School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: zhouw21@lzu.edu.cn; yangmq@lzu.edu.cn; tangjsh20@lzu.edu.cn; bh@lzu.edu.cn).

Juan Wang is with the Department of Psychological Medicine, The Seventh Medical Center of PLA General Hospital, Beijing 100700, China (e-mail: imjuan@sina.com).

Digital Object Identifier 10.1109/TNSRE.2024.3361935

I. Introduction

AUTISM spectrum disorder is a lifelong developmental disorder that can cause social impairment and affect cognition and communication [1]. The Morbidity and Mortality Weekly Report (MMWR) [2] from Centers for Disease Control and Prevention (CDC) released that the overall ASD prevalence rose to 23.0 per 1000 (one in 44) children aged 8 years. Research [3] indicates that timely identification and treatment of autism is critical. Early detection can aid in recognizing children who show signs of the condition, thereby reducing the severity of symptoms and improving their ability to integrate into society. ASD is typically diagnosed using a combination of medical and developmental history, behavioral observations, and standardized assessments. However, there can be subjectivity in the interpretation of behaviors and individual differences in the presentation that may complicate diagnosis. Standardized assessments, such as the Autism Diagnostic Observation Schedule (ADOS) [4] or the Autism Diagnostic Interview-Revised (ADI-R) [5], provide a more structured approach to diagnosis. However, these assessments are time-consuming and may not be widely available, particularly in underserved areas.

Since the etiology and pathogenesis of ASD are not yet clear, there are no specific treatments and reliable clinical diagnostic biomarkers. However, based on the core symptoms of ASD such as social and communication deficits, researchers utilize a variety of behavioral markers. Accurate behavioral phenotype measurements, e.g. facial expression, stereotyped, movement pattern, eye movement, etc., can provide various behavioral markers for ASD diagnosis [6]. For the physiological markers aspect, electrophysiological and neuroimaging studies might provide quite objective indicators from physiological and functional aspects [7]. Zheng et al. [8] reported that brain morphology could provide a basis for discrimination based on the level of abnormal brain structure, which may offer new insights into the neurological underpinnings of atypical visual patterns in autism.

Eye-tracking techniques are widely used in affective computing research and are combined with other modalities to collect multimodal data [9]. It allows a more comprehensive understanding of emotion and its relationship to physiological responses [10]. As a low-cost efficient and non-invasive mea-

surement method to study infants and young children's internal cognitive processes, eye tracking uses relatively objective parameters to avoid possible biases in traditional subjective evaluation [3], [11]. As abnormal eye fixation is one of the behavioral markers and clinical manifestations in early ASD children [12]. Analysis of gaze data has been used to determine the location of important and interesting information within a stimulus [13], [14] and to identify cognitive approaches to task completion [15], [16], which provide the basis for achieving early screening.

Usually, eye tracking is applied to measure subjects' eye gaze data, and the areas of interest (AOIs) based method is used to analyze specific eye movement behaviors in different regions. It defines either priori or posteriori boundaries in the stimulus [17]. The traditional priori AOIs approach, which is based on a top-down strategy, defines the boundaries of AOIs according to experience and the semantic parsing of stimulus, then calculates statistical characteristics within different AOIs [18], [19]. This approach relies on precise and promising manual annotation but lacks a comprehensive examination of visual objects. It only retains data in the region of interest for analysis, and discards fixations that are not within anyone AOI. What is more, it is controversial that the AOI size and location across studies [20]. While a posteriori AOIs method, which is based on a bottom-up strategy, analyzes the eye movement behaviors on the whole stimulus according to the distribution of eye gaze data, provides more objective results by avoiding determining AOIs boundaries [21].

Some studies use a combination of methods for analysis. In [22], complementary data analysis techniques are used to examine the strategies of human face processing by both children with autism and TD children. Three methods are used: the AOI method which defined five regions covering the entire face, the Data-driven method which used the MAP Matlab toolbox to create heat maps, and the Saccade Path method which indicated the frequency of transferring from one interest region to another. This work uses these three techniques in concert, shedding light on potential similarities and differences in face-scanning patterns between ASD and TD children that would not have been observed using the AOI method alone.

Numerous methods have been developed to identify ASD based on eye movements. In recent years, some studies have started to focus on the spatio-temporal model of combining fixations and saccades, known as scanpath. Scanpath has shown high potential in the medical field for diagnosis and treatment, as it has been used as a useful tool for identifying people with schizophrenia [23] and ASD [24], [25]. Analysis of fixation sequences can reveal the cognitive strategies that drive eye movements and provide useful clues for diagnosing whether a subject has ASD or not.

Scanpaths are visual representations that describe eye movement dynamics by characterizing the location and duration of gaze sequences on stimuli [26]. However, there are relatively few studies on scanpath in the field of ASD that have been published so far. Startsev and Dorr [27] combined eye movement statistics, saliency-based, and face-based features to automatically differentiate scanpaths belonging to ASD subjects and control groups. However, the performance of the model in classifying between the training set and the test set is not consistent.

Li et al. [28] make the first attempt to use deep learning techniques to diagnose ASD children in raw video data. They utilize a tracking-learning-detection algorithm to calculate the gaze trajectory of each video and divide it into two components, angle and length. Finally, a three-layer LSTM network is used for classification, and the results show that the LSTM network outperformed traditional machine learning methods. In another study, Carette et al. [29] aim to learn the eye-tracking pattern of ASD by transforming scanpaths into a visual representation. Diagnosis is approached as an image classification task, and neural network models are used to achieve promising classification accuracy on a limited dataset.

In this paper, we propose a digital phenotyping method for studying atypical gaze patterns in children with ASD by analyzing spatio-temporal properties and dynamic changes in scanpaths. Unlike earlier works, our method relies solely on gaze sequences from eye-tracking recordings and takes into account the spatio-temporal dynamics of gaze behaviors. We generate scanpaths and represent them using three components that characterize spatial coordinates, attention duration, and eye movement transfers, along with their dynamics. Our qualitative and quantitative results reveal a significant difference in attention span between children with ASD and TD children. ASD children exhibit greater individual variability in their visual patterns compared to normal children, who tend to develop more similar patterns. Additionally, we employ a classification model with an LSTM network based on scanpaths and achieve a remarkable classification accuracy of 97%.

TABLE I
INFORMATION ABOUT THE PARTICIPANTS

Demographics	ASD	TD
Participants	54	42
Included	54	41
Gender (F,M)	(7, 47)	(23, 18)
Age (m±SD)	5.5±2.29	8.78±3.43

II. DATA COLLECTION

A. Participants

This study involves a group of participants with ASD and a control group of TD. Data were collected from a total of 96 participants (54 ASD and 42 TD) and 95 participants were retained (one in the TD group lacked demographic information). Details about demographic characteristics are shown in Table I. The Ethics Committee of Sichuan Guangyuan Mental Health Center approved the informed consent and study design for the quantitative assessment study of children's autism in accordance with the Declaration of Helsinki, the ethical guidelines of the World Medical Association. All the subjects' guardians agreed that the subjects would participate in this study, signed the informed consent form, and filled in the demographic questionnaires.

1) Inclusion and Exclusion Criteria: The inclusion conditions for the ASD are as follows: (1) participants in the ASD group

have a formal diagnosis of autism and all participants meet the DSM-5 diagnostic criteria; (2) the age range is 12 months to 12 years old, both male and female; (3) IQ is greater than 70. The exclusion criteria for ASD consist of (a) a history of organic brain disease or craniocerebral trauma; (b) genetic or metabolic disease, such as Rett's syndrome; (c) other mental developmental disorders, like attention deficit hyperactivity disorder (ADHD), heller syndrome, etc. (d) Red-green color blindness, diagnosed using a color blindness map. (e) other conditions deemed by the researchers to be inappropriate for participation in this study. Participants in the TD group consist of TD children who must meet (2) and (3) inclusion conditions, and all the exclusion criteria as listed for the ASD group.

B. Materials

Wan et al. [13] proposed that to make it easier for children to participate in experiments, the development of a short and informative paradigm is essential for eye tracking, and their work has shown that a short video clip can provide enough information to distinguish between ASD and TD children. Various types of stimuli are employed for eye-tracking experiments. For example, static faces, stationary moving objects, and other joint attention stimuli. The different stimulus types are viewed as different paradigms, with transitions between them using an animated clip, which may be particularly appealing to children. An eye-tracking experiment usually takes up to 4 minutes. The content and length of the images and videos varied to analyze different aspects of eye behaviors.

We use the human face images experiment paradigm in our subsequent analysis. Due to its superior performance in early ASD research, it has become the most commonly used experimental paradigm for eye tracking in ASD detection [30], [31].

The human face stimulus set consists of a 34-second video segment that includes six images of a woman's face expressing different emotions from the Chinese facial affective picture system (CFAPS) [32]. Each image consists of a frontal woman's face above the neck with a solid black background and lasts for 5 seconds.

C. Experimental Setup

The eye-tracking experiment consisted of a set of image and video scenes specifically designed to stimulate the gaze on different screen parts. The experiment is conducted in a quiet room with only the researchers, the subjects, and their guardians. The distance between the participant and the eye tracker is approximately $60cm$. To record the children's gaze behaviors, we employ a 90HZ Tobii Eye Tracker 4C video-based eye tracker below a 23-inch display that the LCD monitor with a resolution set to 1920×1080 pixels. An RGB camera is set up on the side to record the whole experiment, as shown in Fig. 2.

Before commencing the experiment, participants are required to engage in a preparatory session aimed at acquainting them with the experimental environment and procedures. A five-point calibration is performed using the Tobbi internal program. Calibration is deemed successful when all five points exhibit a strong fit in the computational mapping. The eye-tracker is responsible for capturing the subject's gaze behavior, and the ordinary camera records the subject's viewing process.

III. Method

In this section, we will provide a detailed description of our proposed method, which includes feature extraction of raw eye-tracking data, classification, and similarity analysis, as shown in Fig.1. By characterizing the scanpath that reflects changes in fixations and saccades, we differentiate between ASD and TD children and conduct qualitative and quantitative analyses of differences in continuous visual behavioral patterns.

A. Fixations and Saccades Identification

Velocity-threshold fixation identification (I-VT) [33] algorithm which identifies fixations from saccades based on the fact that fixations have lower velocities than the saccades is the simplest of the identification methods to understand and implement that is widely accepted in eye-tracking protocols. It requires one specification parameter, the velocity threshold that can be computed when the distance from the eye to visual stimuli is known [34]. Karthik et al. [35] achieved a custom implementation of the velocity threshold algorithm for fixation identification. They found that even by considering only points of regard to finding fixations, results obtained are similar to those obtained by running SMI BeGaze. Their work provides us with a potential reference.

The raw eye gaze data are generally expressed by sequences of sampling gaze points. Compared to controls, children with ASD spend less time attending to people and goal-oriented tasks so they have a harder time maintaining sustained attention and fewer fixations on the monitor. The definition of velocity is $v = d/t$ and the sampling rate is provided above. Therefore, only the distance between the two fixation positions is required to calculate v. In our study, sequences with missing values of less than 10 sample points are retained. We calculate the original velocity of each sampling point in the individual scanpath. If the v of the point is below a predefined threshold, then the point will be considered a fixation. Otherwise, it will be considered a saccade and be excluded.

B. DTW and MultiMatch Based Scanpath Similarity

In this paper, we use two different measures of similarity to explore differences in scanpaths between ASD children and TD children. Fig. 3 shows a brief diagram of both approaches.

The Dynamic Time Warping (DTW) [36] method is suitable for calculating the similarity of time series, especially for those with different lengths, which correspond to the characteristics of the scanpath. Each scanpath may have a different length, as the duration of gaze and the frequency of gaze shifts are subjectively controlled. Given two scanpaths $Q =< q_1, q_2, \ldots, q_m >$ and $P =< p_1, p_2, \ldots, p_n >$, where m and n represent the length of the sequence Q and P,

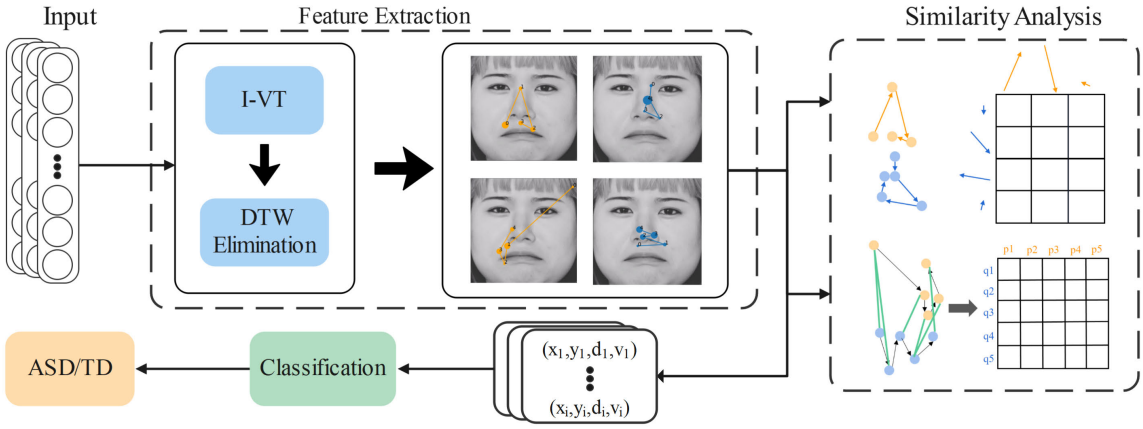
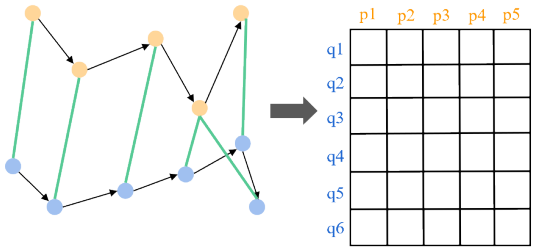
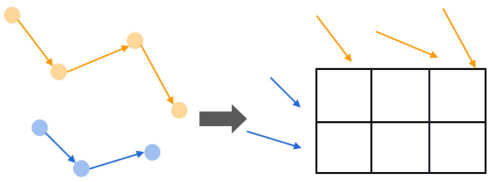


Fig. 1. The framework of the proposed method in this paper. It consists of four steps: (1) velocity calculation; (2) outlier elimination; (3) construction of networks for classification; (4) similarity measurement for analysis based on MultiMatch and DTW.

Fig. 2. The experimental scene layout of this study.



(a) Dynamic Time Warping



(b) MultiMatch

Fig. 3. Schematic diagram of the two similarity methods calculated.

respectively. The DTW distance is recursively computed by the following formula:

$$D(Q_i, P_j) = \delta(q_i, p_j) + \min \begin{cases} D(Q_{i-1}, P_{j-1}) \\ D(Q_{i-1}, P_j) \\ D(Q_i, P_{j-1}) \end{cases} \quad (1)$$

where Q_i, P_j are the subsequences of Q and P, $\delta(q_i, p_j)$ is the Euclidean distance between points q_i and b_j, $D(Q_i, Pj)$ is the DTW distance between Q_i and P_j. Therefore, we obtain a DTW matrix with the DTW distance between any pairwise subsequences of two given scanpaths. Based on this matrix, we can find the path of cost accumulation from the bottom right to the top left, which indicates the optimal alignment between sequences.

On the other hand, we estimate scanpath through similarity, which is evaluated by MultiMatch [37], [38]. MultiMatch is not simply comparing the similarity of fixation points, it also takes into account the sequential relationship between different fixations and is capable of handling the similarity between multiple scanpaths at once. Moreover, MultiMatch provides five different evaluation metrics, making it adaptable to diverse eye-tracking data analysis needs. By splitting a single scanpath into multiple vector segments, the MultiMatch method directly compares these vectors from five aspects that capture the similarity between different characteristics of scanpaths, namely vector similarity, direction similarity, length similarity, position similarity, and duration similarity. The two endpoints of a vector represent consecutive gaze points, and the vector segments represent the saccade behavior between consecutive gaze.

Given two vectors consisting of a pair of fixations and saccades from different scanpaths Q and P, represented by $\mathbf{u}_i$ and $\mathbf{v}_i$. The average of the following measured values will be calculated:

- $|\mathbf{u}_i - \mathbf{v}_i|$ represents the shape difference.
- $\|\mathbf{u}_i - \mathbf{v}_i\|$ represents the length difference.
- The Euclidean distance is used to indicate the distance of the fixations.
- The direction difference is represented by the angle between $\mathbf{u}_i$ and $\mathbf{v}_i$.
- The duration of the fixations indicates the difference in the vector's duration.

The values on each metric are normalized between 0 and 1, the screen diagonal is used as a reference. π is used to normalize direction, and the difference for the duration is normalized against the maximum duration of the two fixations

being compared. Finally, we end up with five similarity scores.

C. MIC Score Based Path and Similarity

The Maximal Information Coefficient (MIC) [39] method can capture a wide range of associations between features and labels and is not limited to specific types. Compared with the embedding method which combined some machine learning models to get the weight coefficient of features, the bias of selecting different machine learning models is avoided. The correlation is obtained by calculating "the estimation of the mutual information between each feature and the label".

To explore the interpretability of the scanpaths, we add the extra length attribute for each fixation for analysis. Given the features $f_i = (x_i, y_i, d_i, v_i, l_i)$ and labels s_i. To encode the scatter plot consisting of a label and each feature of f into a grid, the mutual information between features and labels is calculated by joint distribution, as follows:

$$I(f,s) \approx I(F,S) = \sum_f \sum_s p(f,s) log_2 \frac{p(f,s)}{p(f)p(s)} \quad (2)$$

And normalized the value into the range [0, 1] (0 means the two variables are independent, and 1 means the two variables are completely correlated).

$$mic(f,s) = \frac{max_{G \in G(f,s)} I(F(G), S(G))}{log_2 min(f,s)} \quad (3)$$

Notice that we calculate MIC values in turn for the features in f. And $G(f,s)$ is the set of two-dimensional grids of size $f \times s$.

D. Features

Outlier scanpaths will introduce noise to the distribution trend of the population over time and space. We observe the dispersion of scanpaths using five statistical measures, including the minimum, lower quartile, median, upper quartile, and maximum. Based on the DTW distance, values that deviate from the upper or lower quartile by 1.5 times the interquartile range in either of the two groups are considered outlier scanpaths and removed.

In the first step, the average DTW distance is calculated between each scanpath and all remaining scanpaths in the same group. This value is then used to calculate statistical characteristics for outlier elimination. It is worth noting that every time excluded one scanpath, the average DTW distance of the remaining scanpaths will be changed and recalculated accordingly.

In scanpaths, we extract four features for each fixation, $f_i = (x_i, y_i, d_i, v_i)$ namely x_pos, y_pos, $duration$, and $velocity$. Where x_pos and y_pos represent the coordinates of the participant's gaze on the monitor, $duration$ is the level of sustained attention, and $velocity$ means the speed at which the current fixation transferred to the next. When one fixation consists of multiple sampling points whose velocity is less than the velocity threshold, x_pos, y_pos, and $velocity$ are the average of these points, while $duration$ is the sum of the duration of each one. These features reflect the current location of interest, duration, and speed of shift of attention.

When the number of positive and negative samples is not equal, or even the gap is large, a larger proportion of samples of a certain category will account for most of the loss values. Therefore, the model learns less for the smaller sample size category, resulting in a worse generalization of the model. In our study, after the pre-process, the number of scanpaths in the ASD group is found to be significantly lower than that in the TD group. In order to balance the two groups, we carry out the following procedure: Each face stimulus image is treated as a distinct experimental paradigm, and the scanpaths that have been processed as described above are considered valid data for each paradigm. Then, all scanpaths within each paradigm are sorted according to the DTW distance, and an equal number of samples are retained for the ASD and TD groups to eliminate the bias that may have been introduced by the imbalance of samples in the classification model.

E. Classification

LSTM can effectively express and convey information in a long time series without forgetting useful information long ago, so it is very suitable for sequence. Based on this characteristic of LSTM, this study uses LSTM and different variant networks of LSTM for training and classifying ASD and TD. After feature extraction, we fed four different combinations of features into the following networks: (1) GRU. (2) 2-LSTM. (3) Bidirectional LSTM (BiLSTM).

To achieve the training of unequal-length sequences, we use the Mask mechanism for processing before inputting the model. Given the maximum length of scanpaths, the Mask mechanism would pad other sequences up to that length and then remove these parameters during training. For classification, each layer of the LSTM has 128 hidden units, and only the outputs of all hidden units in the last layer are taken and then fed into the fully connected layer, which uses a softmax regression layer to predict the probability of each class for binary classification. All experiments are performed on a batch size 64 with a cross-entropy loss function and a learning rate of 0.0001.

For performance and evaluation, we use a 5-fold cross-validation for classification. In each fold, 70% of the data is used as a training set, 10% as a validation set, and the remaining is used for testing. We repeat this process 5 times and finally average the evaluation results for each classification.

In addition, traditional machine learning methods are also considered for comparison. The features are fed into a traditional machine classification model to verify whether the neural network model has better performance in the scanpaths classification task. Random forest (RF) can provide information about the relationship between variables and labels, while XGBoost uses multiple weak classifiers to integrate a strong classifier, which can control the complexity of the model and prevent overfitting. We also take a 5-fold cross-validation for classification.

To evaluate the performance of the binary classification, we employ accuracy (Acc.), recall (Rec.), and f1_score (F1.) as our evaluation indicators. Details of the classification results

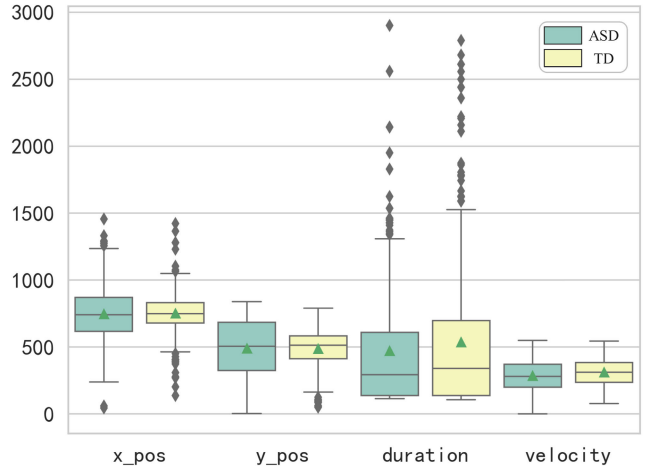


Fig. 4. Comparison of scanpaths' features. A boxplot is drawn on the four features. It can be found that the difference between position and duration is significant, but the velocity distribution of the two groups is similar.

are shown in Table III. We can find that no matter what kind of classifier, when the feature set contained duration (see row2 and row4), the accuracy is higher. And better classification results are achieved on LSTM and GRU networks. We speculate that complex network models (2)-LSTM and Bi-LSTM) bring additional redundancy parameter information due to a more complex network structure. As for machine learning models (RF and XGBoost), the classification accuracy is inferior to that of neural network models due to the lack of time series information.

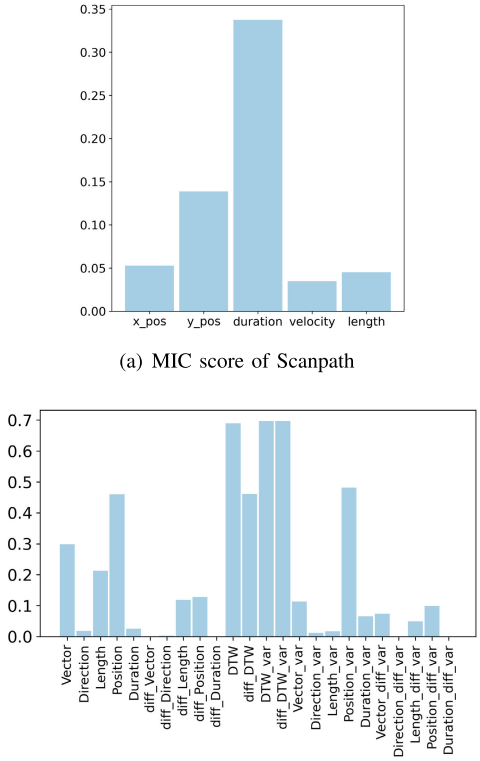
IV. Results

We aim to reveal the differences in eye movement behavior between individuals with and without ASD by exploring similarities within and between scanpaths. This approach enables us to capture both individual variability and group differences, and we use the MIC to offer potential explanations. Our findings suggest that accounting for continuous visual behavior can yield more valuable information than relying solely on discrete fixation points, and both the classification and analysis results confirm that duration is the most notable difference between the two groups.

A. Scanpath Measurements

For the scanpaths, we obtain the features of five dimensions, $f_i = (x_i, y_i, d_i, v_i, l_i)$, including the four features of the input neural network and the newly added length feature. Taken these features as the input of the MIC method to get the correlation coefficient. The discrete distribution of features is shown in Fig. 4. From the results, we can observe that $duration$ is associated with the label the most, followed by the y. It indicates that these two features are most associated with classes.

1) Scanpath Position: Many studies have identified atypical gaze behaviors in children with autism, such as eye gaze avoidance, using a fixation-based AOI approach. However,



(a) MIC score of Scanpath

(b) MIC score of Scanpath Similarity

Fig. 5. MIC score calculation from the scanpath and the similarity of scanpaths. A significant difference in position can be observed from both results, and the DTW takes Euclidean distance as elements are more sensitive to changes on the Y-axis to capture this otherness. For the duration, the MultiMatch segmenting scan path shows no distinction between within and between ASD, while the DTW does not include gaze duration in its calculations.

in our study, the coordinates of fixation points are divided into x and y for analysis. As shown in Fig. 4, we can find that the gazing distribution of ASD is more discrete in the vertical direction than in the horizontal direction compared to the TD group. This suggests that children with ASD tend to have more up-down shifted attention when looking at faces.

2) Scanpath Duration: When exploring a visual scene, we can locate the visual information of a certain part of the image through fixation. The duration reflects the accumulated attention of the focus area and is related to the processing of visuals. According to the MIC results, duration has the strongest correlation with groups, which indicates a significant difference between the two groups. As shown in Fig. 4 and Fig. 5(a), we can observe that the duration of gaze in children with ASD is generally shorter than that of TD children, whether in terms of the mean or median values. It reflects a gaze pattern of TD children, which is that for most given images, they tend to focus on certain regions and maintain attention for a period of time.

3) Scanpath Velocity: Velocity reflects the eye movement behavior of moving from the current fixation to the next, during which little or no valid visual information is acquired. We can notice that velocity has the least correlation with groups, which is consistent with the classification results. As shown in Table III, row 2 is significantly improved compared with

TABLE II
AVERAGE OF MULTIMATCH AND DTW SIMILARITY MEASUREMENT

Classes	Vector	Direction	Length	Position	Duration	DTW
ASD:ASD	0.92931	0.64187	0.93518	0.85373	0.44679	403.225
TD:TD	0.94458	0.66127	0.94621	0.90667	0.43141	132.559
ASD:TD	0.93464	0.63761	0.93623	0.87283	0.43784	305.586
TD:ASD	0.94909	0.64759	0.95086	0.88641	0.44519	305.586

TABLE III
COMPARISON BETWEEN DIFFERENT CLASSIFIERS

Features	Metrics	LSTM	GRU	2-LSTM	Bi-LSTM	RFC	XGBoost
(x, y)	Acc.(%)	94.00	91.00	85.00	84.50	71.23	70.99
	Rec.(%)	95.18	94.25	71.16	69.76	73.11	73.33
	F1.(%)	94.11	91.15	83.05	81.98	73.85	73.74
(x, y, d)	Acc.(%)	97.00	92.00	92.00	90.00	80.80	80.25
	Rec.(%)	100.00	92.66	83.89	79.92	80.22	78.89
	F1.(%)	97.08	91.61	91.19	88.67	82.28	81.61
(x, y, v)	Acc.(%)	87.00	84.50	88.00	91.00	70.25	65.43
	Rec.(%)	74.76	74.62	75.21	89.94	71.22	64.44
	F1.(%)	85.42	82.98	85.69	90.85	72.67	67.44
(x, y, d, v)	Acc.(%)	89.00	92.50	90.00	87.50	80.10	77.16
	Rec.(%)	77.24	100.00	87.62	83.60	77.44	75.56
	F1.(%)	86.85	93.27	90.01	87.29	81.90	78.61

row 1 on all models, while row 3 has no such performance. So we suggest that velocity does not provide additional useful information. In other words, the process of transfer attention can not provide visual pattern distinctions, but the continuous attention process reflects atypical behavior in individuals with ASD.

4) Scanpath Length: The length of scanpaths reflects the frequency of attention shift. Li et al. [40] analyzed the length of scanpaths in two publicly available datasets, OSIE and MIT1003, and found that the distribution of length is similar to normal distribution. This suggests that when viewing a given stimulus, most subjects exhibit similar gaze distribution patterns, but a small number of individuals exhibit infrequent shifts or rapid changes of attention. Since gaze behavior is subjectively controlled, the length of scanpaths may vary among subjects, but this bell-shaped distribution characteristic is shared. In the data we collected, both ASD and TD children maintained this distribution of scanpath length.

B. Similarity

It is important to take into account the required quantification format when choosing a scanpath comparison method [41]. The MultiMatch method employs a direct quantification, which enables the direct comparison of eye movement behaviors across scanpaths. Using the MultiMatch method, we obtain similar values on the five features of scanpaths, and the values on each metric are normalized between 0 and 1. The larger the scores, the higher the similarity.

In addition, we use the DTW to calculate the similarity between the two groups. The DTW uses a global optimization strategy to consider the best alignment between the two sequences without other processing. We expect that scanpaths of the same class are as similar as possible, while different classes are as far apart as possible. And a smaller DTW distance means a better similarity.

1) Overall Scanpath Similarity: We compare the scanpath similarity with all scanpaths in the same classes and all in controls using both algorithms and compute the average (Table II) scores. From the results of MIC as shown in Fig. 5(b), we can observe that the correlation between the DTW and position is much higher than other characteristics, which corresponds to the results in Table II. The mean values of the DTW in the TD group are much smaller than that in the ASD. However, it is not easy to distinguish the difference between the two groups in the score of the MulitiMatch method. This may be because MultiMatch simplifies or quantifies the scanpath before comparison, whereas DTW directly uses the Euclidean distance between fixation points as the element of the cost matrix.

C. Comparison on Open Dataset

To validate the generalizability of our method, we performed a comparison on an open dataset of eye movements of children with ASD [42]. It consists of 300 natural scene images and corresponding eye movement data collected from 14 children with ASD and 14 TD children. The Saliency4ASD Grand Challenge "Saliency4ASD: Visual attention modeling for Autism Spectrum Disorder" was organized (presented at ICME 2019), which is based on publicly available datasets mentioned above to align the visual attention modeling community around the application of characterizing and diagnosing ASD.

Considering that this public dataset is the latest and most similar to our experimental data, we apply our method to this dataset and compare it with all methods published at ICME. We conducted two separate experiments, one using all the eye movement data from 300 natural scene images for training and classification, and the other selecting the eye movement data corresponding to the images that contain human themes. It is a subset of all the images. We select the LSTM model

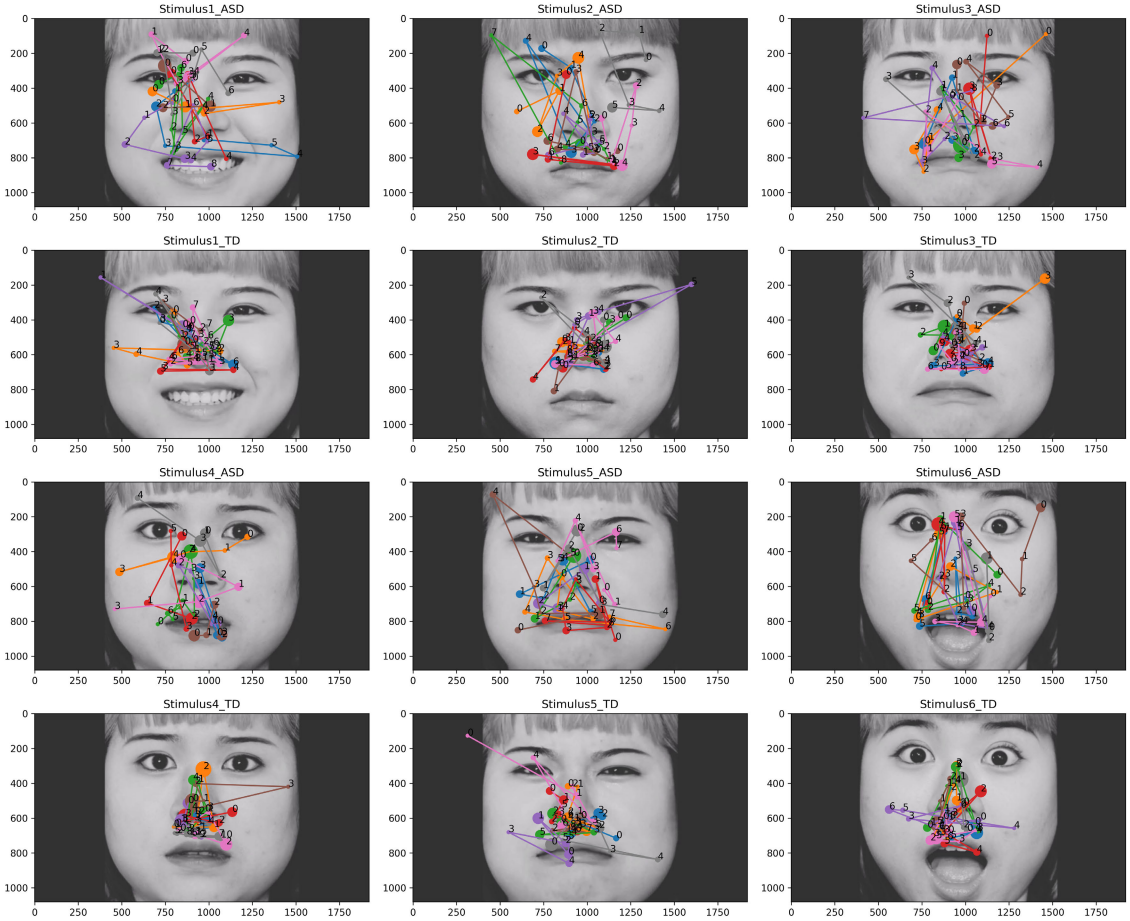


Fig. 6. Visualization of scanpath. Different colors represent the scanpaths of different individuals. On each scanpath, the size of the circle indicates the duration of each fixation point, and the number identifies the order in which the fixation points are observed.

with the most superior performance in our experiments as the classifier. The experimental procedure remains consistent with the above, and only the model parameters epoch and learning rate are adjusted for different data volumes.

As shown in Table IV, Startsev and Dorr [27] in their study used AUC scores as an evaluation metric and extracted the basic fixation statistics of each scanpath as the features, which ultimately achieved 75% AUC. Their further analysis showed that eye movement data containing multiple face images had a higher discriminative ability (76.9% AUC on average). Wu et al. [43] employed the methods of classification using images with gaze and synthetic saccade, respectively, and achieved better results on the latter with an accuracy of 65.41%. Tao and Shyu [44] proposed an SP-ASDNET network that combines CNN and LSTM and the best classification result accuracy is 74.22%. Using our method, the average accuracy of cross-validation on the full dataset is 83.93%, while on the dataset containing only faces, the average accuracy achieved is 87.95%. The experimental results show that our method has better classification performance than others and can capture discriminative features of eye movement data better on human face image paradigms. Furthermore, the application across datasets indicates the effectiveness and generalizability of the proposed method for ASD classification in this paper.

TABLE IV
COMPARISON OF DIFFERENT CLASSIFICATION METHODS ON SALIENCY4ASD DATASET

Methods	Database	Acc.(%)	AUC(%)
SP-ASDNET [44]	Full Dataset	74.22	/
Synthetic Saccade [43]	Full Dataset	65.41	/
Image-based [43]	Full Dataset	61.62	/
Fixation statistics [27]	Full Dataset	/	75
	Human Face Dataset	/	76.9
Our Merthod	Full Dataset	83.93	83.89
	Human Face Dataset	87.95	88.38

D. Visualization of Scanpaths

Scanpath visualization shows the distribution of visual attention when subjects observe a given image, aiding in the analysis and comprehension of their visual attention preferences. The results of scanpath visualization for each facial stimulus image are displayed in Fig. 6. Overall, children with ASD tend to have a gaze distribution that roams over the entire image, while TD children tend to concentrate on the central part of the image. Specifically, ASD children are inclined to focus on the left half of the face (below the left eye), while the TD group tends to focus on the middle part of the face (the nose area). Buchan et al. [45] utilized the AOI method

to assess individual preferences and found that some subjects prefer to gaze at the left half of the face, whereas others prefer to gaze at the right half. This reflects the asymmetry of facial expressions, wherein the right side of the face is better at expressing emotions than the left side [46], which may have a potential correlation with the avoidance of the right half of the face by ASD children.

V. Discussion

Eye tracking studies have the potential to provide a unique and intermediate-level description of autism, linking underlying neurocognitive networks to the upper level of everyday functioning and dysfunction [47]. And it can reveal how people with ASD process social visual information.

Unlike prior research that analyzes gaze behavior using statistical features across distinct facial regions, this study examines fixation locations and discovers that children with ASD exhibit larger vertical gaze shift amplitudes than horizontal ones. This finding is consistent with the results of Wegner-Clemens et al. [48]. They find that ASD children and TD children exhibit significant individual differences in their tendency to gaze at the eyes or mouth, confirming that the distribution of gaze on the upper and lower face is the largest difference between the two groups. The next largest difference is the distribution of gaze on the left and right halves of the face. It may explain why the distribution of gaze in the vertical direction is more dispersed than in the horizontal.

By examining the degree of association between features and labels, we discover that duration is the most significant feature for distinguishing between ASD and TD children. In contrast, velocity, which characterizes the process of gaze shift, does not provide additional information for classification. This suggests that the ability to sustain attention is the primary difference between the two groups. Specifically, during the process of viewing facial images, ASD children exhibit shorter fixation duration and more frequent shifts, while TD children tend to generally maintain their visual attention.

We further analyze scanpath similarity from a visual behavior perspective. As gaze behavior is a continuous process, individual differences in visual behavior can be exposed by the overall shape of the scanpath and the order in which it unfolds. In this study, we utilize two different similarity measures. The MultiMatch method represents a single scanpath as multiple segments of vectors and calculates alignment by optimizing vector differences between scanpaths. However, this method may reduce sensitivity to minor temporal or spatial changes. The DTW method aims to make the shapes of two sets of sequences as similar as possible by warping them. Through the above analysis, we can observe that the visual distribution of ASD children differs more in the vertical direction, and the DTW method, which employs Euclidean distance as an element, is more sensitive to distance in the Y direction. Finally, we compare multiple neural networks and machine learning methods to demonstrate the effectiveness of LSTM in continuous visual behavior.

Our study also has some limitations to consider. Firstly, small-scale datasets are currently a common problem in the ASD research field. To obtain a sufficient amount of valid data, we preserve valid scanpaths in any paradigm. However, this may lead to a problem that an ASD child has valid data in one paradigm but not in others, resulting in different subjects having different sample sizes. Secondly, the DTW method is commonly applied in the field of speech recognition, which only considers the alignment of gaze points in scanpaths without considering duration. In future work, we will focus on improving this algorithm by incorporating duration as a weight in the alignment process.

Finally, there are also some limitations of the eye-tracking technology. Visual exploration is a highly specific and personalized process. Therefore, the eye tracker needs to be calibrated before use by each subject to ensure accuracy. If the calibration process is not accurate, the data may be offset during subsequent acquisition. Data offset is one of the sources of noise. Data offset may also occur when the position of the eye relative to the camera is shifted, or when the pupil size changes. In addition, multiple calibrations can create a burden on the time and labor costs of the experiment.

In our study, children between the ages of 3 and 12 are included. This contributes to facilitating that the participating children can understand the experimental task to ensure their cooperation. For some children, maintaining sufficient attention and cooperation to complete calibration can be challenging. Data integrity checks and denoising also lead to a reduction in the size of the dataset. In addition, current eye-tracking methods require special eye-tracking equipment, which is not inexpensive enough, which limits the widespread use of the method in autism screening, especially in extremely underdeveloped regions.

VI. Conclusion

In this study, we focus on the scanpaths which are seen as visual representations that describe eye movement dynamics on stimuli and there are rarely works specifically for it. We propose a scanpath-based ASD detection method that aims to learn the atypical visual patterns of ASD. This digital phenotyping method provides qualitative and quantitative results. The classification results show that LSTM outperforms traditional machine learning methods. In terms of intra- and inter-group similarity in scanpaths, it reveals that ASD children exhibit greater individual variability from three dimensions in visual patterns, which are spatial coordinates, attention duration and eye movement transfers.

References

[1] *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., Amer. Psychiatric Assoc., Washington, DC, USA, 2013, pp. 591–643.

[2] M. J. Maenner et al., "Prevalence and characteristics of autism spectrum disorder among children aged 8 years–Autism and developmental disabilities monitoring network, 11 sites, United States, 2018," *MMWR Surveill. Summaries*, vol. 70, no. 11, pp. 1–16, 2018.

[3] L. Zwaigenbaum and M. Penner, "Autism spectrum disorder: Advances in diagnosis and evaluation," *BMJ*, vol. 361, May 2018, Art. no. k1674.

[4] C. Lord et al., "The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Develop. Disorders*, vol. 30, pp. 205–223, Jun. 2000.

[5] M. Rutter, A. Le Couteur, and C. Lord, *Autism Diagnostic Interview Revised*, vol. 29. Los Angeles, CA, USA: Western Psychol. Services, 2003, p. 30.

[6] R. A. J. de Belen, T. Bednarz, A. Sowmya, and D. Del Favero, "Computer vision in autism spectrum disorder research: A systematic review of published studies from 2009 to 2019," *Translational Psychiatry*, vol. 10, no. 1, p. 333, Sep. 2020.

[7] C. Lord et al., "Autism spectrum disorder," *Nature Rev. Disease Primers*, vol. 6, no. 1, pp. 1–23, 2020.

[8] W. Zheng et al., "Multi-feature based network revealing the structural abnormalities in autism spectrum disorder," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 732–742, Jul. 2021.

[9] M. Yang, Y. Gao, L. Tang, J. Hou, and B. Hu, "Wearable eye-tracking system for synchronized multimodal data acquisition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 14, 2023, doi: 10.1109/TCSVT.2023.3332814.

[10] M. Yang, Y. Wu, Y. Tao, X. Hu, and B. Hu, "Trial selection tensor canonical correlation analysis (TSTCCA) for depression recognition with facial expression and pupil diameter," *IEEE J. Biomed. Health Informat.*, early access, Oct. 5, 2024, doi: 10.1109/JBHI.2023.3322271.

[11] A. Klin, "Biomarkers in autism spectrum disorder: Challenges, advances, and the need for biomarkers of relevance to public health," *Focus*, vol. 16, no. 2, pp. 135–142, Apr. 2018.

[12] 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, pp. 427–431, Dec. 2013.

[13] G. Wan et al., "Applying eye tracking to identify autism spectrum disorder in children," *J. Autism Develop. Disorders*, vol. 49, no. 1, pp. 209–215, Jan. 2019.

[14] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, "Detecting high-functioning autism in adults using eye tracking and machine learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1254–1261, Jun. 2020.

[15] Z. Zhao, H. Tang, X. Zhang, X. Qu, X. Hu, and J. Lu, "Classification of children with autism and typical development using eye-tracking data from face-to-face conversations: Machine learning model development and performance evaluation," *J. Med. Internet Res.*, vol. 23, no. 8, Aug. 2021, Art. no. e29328.

[16] C. Tang et al., "Automatic identification of high-risk autism spectrum disorder: A feasibility study using video and audio data under the still-face paradigm," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2401–2410, Nov. 2020.

[17] M. E. Minissi, I. A. C. Giglioli, F. Mantovani, and M. A. Raya, "Assessment of the autism spectrum disorder based on machine learning and social visual attention: A systematic review," *J. Autism Develop. Disorders*, vol. 52, no. 5, pp. 2187–2202, May 2022.

[18] O. Kardan, M. G. Berman, G. Yourganov, J. Schmidt, and J. M. Henderson, "Classifying mental states from eye movements during scene viewing," *J. Exp. Psychol., Human Perception Perform.*, vol. 41, no. 6, pp. 1502–1514, Dec. 2015.

[19] J. Lao, S. Miellet, C. Pernet, N. Sokhn, and R. Caldara, "IMap4: An open source toolbox for the statistical fixation mapping of eye movement data with linear mixed modeling," *Behav. Res. Methods*, vol. 49, no. 2, pp. 559–575, Apr. 2017.

[20] J. Irwin, T. Avery, L. Brancazio, J. Turcios, K. Ryherd, and N. Landi, "Electrophysiological indices of audiovisual speech perception: Beyond the McGurk effect and speech in noise," *Multisensory Res.*, vol. 31, nos. 1–2, pp. 39–56, 2018.

[21] F. Cilia, A. Aubry, B. Le Driant, B. Bourdin, and L. Vandromme, "Visual exploration of dynamic or static joint attention bids in children with autism syndrome disorder," *Frontiers Psychol.*, vol. 10, p. 2187, Oct. 2019.

[22] L. Yi et al., "Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data," *J. Vis.*, vol. 13, no. 10, p. 5, Aug. 2013.

[23] C. M. Loughland, L. M. Williams, and E. Gordon, "Visual scanpaths to positive and negative facial emotions in an outpatient schizophrenia sample," *Schizophrenia Res.*, vol. 55, nos. 1–2, pp. 159–170, May 2002.

[24] K. A. Pelphrey, N. J. Sasson, J. S. Reznick, G. Paul, B. D. Goldman, and J. Piven, "Visual scanning of faces in autism," *J. Autism Develop. Disorders*, vol. 32, no. 4, pp. 249–261, 2002.

[25] K. Horley, L. M. Williams, C. Gonsalvez, and E. Gordon, "Face to face: Visual scanpath evidence for abnormal processing of facial expressions in social phobia," *Psychiatry Res.*, vol. 127, nos. 1–2, pp. 43–53, Jun. 2004.

[26] J. H. Goldberg and J. I. Helfman, "Visual scanpath representation," in *Proc. Symp. Eye-Tracking Res. Appl. (ETRA)*, 2010, pp. 203–210.

[27] M. Startsev and M. Dorr, "Classifying autism spectrum disorder based on scanpaths and saliency," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 633–636.

[28] J. Li, Y. Zhong, J. Han, G. Ouyang, X. Li, and H. Liu, "Classifying ASD children with LSTM based on raw videos," *Neurocomputing*, vol. 390, pp. 226–238, May 2020.

[29] R. Carette, M. Elbattah, F. Cilia, G. Dequen, J.-L. Guérin, and J. Bosche, "Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths," in *Proc. 12th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2019, pp. 103–112.

[30] T. X. Fujisawa, S. Tanaka, D. N. Saito, H. Kosaka, and A. Tomoda, "Visual attention for social information and salivary oxytocin levels in preschool children with autism spectrum disorders: An eye-tracking study," *Frontiers Neurosci.*, vol. 8, p. 295, Sep. 2014.

[31] F. Shic, S. Macari, and K. Chawarska, "Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder," *Biol. Psychiatry*, vol. 75, no. 3, pp. 231–237, Feb. 2014.

[32] X. Gong, Y.-X. Huang, Y. Wang, and Y.-J. Luo, "Revision of the Chinese facial affective picture system," *Chin. Mental Health J.*, vol. 25, no. 1, pp. 40–46, 2011.

[33] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, 2000, pp. 71–78.

[34] A. F. Fuchs, "The saccade system," in *The Control of Eye Movements*. Amsterdam, The Netherlands: Elsevier, 1971, pp. 343–362.

[35] G. Karthik, J. Amudha, and C. Jyotsna, "A custom implementation of the velocity threshold algorithm for fixation identification," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 488–492.

[36] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Feb. 1978.

[37] H. Jarodzka, K. Holmqvist, and M. Nyström, "A vector-based, multi-dimensional scanpath similarity measure," in *Proc. Symp. Eye-Tracking Res. Appl. (ETRA)*, 2010, pp. 211–218.

[38] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, "It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach," *Behav. Res. Methods*, vol. 44, no. 4, pp. 1079–1100, Dec. 2012.

[39] B. Robidoux, "Maximal information coefficient: An introduction to information theory," *Predictive Anal. Futurism*, no. 15, pp. 1–44, Jun. 2017.

[40] A. Li and Z. Chen, "Representative scanpath identification for group viewing pattern analysis," *J. Eye Movement Res.*, vol. 11, no. 6, Nov. 2018, doi: 10.16910/jemr.11.6.5.

[41] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof, "A comparison of scanpath comparison methods," *Behav. Res. Methods*, vol. 47, no. 4, pp. 1377–1392, Dec. 2015.

[42] H. Duan et al., "A dataset of eye movements for the children with autism spectrum disorder," in *Proc. 10th ACM Multimedia Syst. Conf.*, Jun. 2019, pp. 255–260.

[43] C. Wu, S. Liaqat, S.-C. Cheung, C.-N. Chuah, and S. Ozonoff, "Predicting autism diagnosis using image with fixations and synthetic saccade patterns," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 647–650.

[44] Y. Tao and M.-L. Shyu, "SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 641–646.

[45] J. N. Buchan, M. Paré, and K. G. Munhall, "The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception," *Brain Res.*, vol. 1242, pp. 162–171, Nov. 2008.

[46] W. R. Powell and J. A. Schirillo, "Asymmetrical facial expressions in portraits and hemispheric laterality: A literature review," *Laterality*, vol. 14, no. 6, pp. 545–572, Nov. 2009.

[47] T. Falck-Ytter, S. Bölte, and G. Gredebäck, "Eye tracking in early autism research," *J. Neurodevelopmental Disorders*, vol. 5, no. 1, pp. 1–13, Dec. 2013.

[48] K. Wegner-Clemens, J. Rennig, J. F. Magnotti, and M. S. Beauchamp, "Using principal component analysis to characterize eye movement fixation patterns during face viewing," *J. Vis.*, vol. 19, no. 13, p. 2, Nov. 2019.