Signal Processing: Image Communication 94 (2021) 116198

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Predicting ASD diagnosis in children with synthetic and image-based eye gaze data[☆]

Sidrah Liaqat[a,1], Chongruo Wu[b,1], Prashanth Reddy Duggirala[b], Sen-ching Samson Cheung[a,b,*], Chen-Nee Chuah[b], Sally Ozonoff[b], Gregory Young[b]

[a] *University of Kentucky, United States of America*
[b] *University of California, Davis, United States of America*

ARTICLE INFO

Keywords:
Autism spectrum disorders
Eye gaze data
Deep learning

ABSTRACT

As early intervention is highly effective for young children with autism spectrum disorder (ASD), it is imperative to make accurate diagnosis as early as possible. ASD has often been associated with atypical visual attention and eye gaze data can be collected at a very early age. An automatic screening tool based on eye gaze data that could identify ASD risk offers the opportunity for intervention before the full set of symptoms is present. In this paper, we propose two machine learning methods, synthetic saccade approach and image based approach, to automatically classify ASD given children's eye gaze data collected from free-viewing tasks of natural images. The first approach uses a generative model of synthetic saccade patterns to represent the baseline scan-path from a typical non-ASD individual and combines it with the real scan-path as well as other auxiliary data as inputs to a deep learning classifier. The second approach adopts a more holistic image-based approach by feeding the input image and a sequence of fixation maps into a convolutional or recurrent neural network. Using a publicly-accessible collection of children's gaze data, our experiments indicate that the ASD prediction accuracy reaches 67.23% accuracy on the validation dataset and 62.13% accuracy on the test dataset.

1. Introduction

Autism spectrum disorder (ASD) is defined by the deficits in social and communication development and the presence of stereotypical behaviors. The natural course of ASD involves symptom onset in the first three years of life. Multiple studies have demonstrated that differences between children who will later receive an ASD diagnosis and those who develop typically often emerge well before the second birthday [1,2]. These differences, which include limited eye contact, shared affect, and joint attention, have been demonstrated using multiple methodologies. Despite this promise for early identification, the mean age of ASD diagnoses in the United States is still over 4 years [3], with less than 25% made before age 3 [4], squandering years of potential intervention when the brain is most plastic. As such, there is an urgent need in developing robust and easy-to-use ASD screening tools for infants and toddlers.

ASD has often been associated with atypical visual attention, sometimes emerged even before the onset of the disorder [5]. One promising direction for early detection is to consider ASD prediction as a classification problem and use machine learning (ML) techniques to differentiate visual gaze patterns between individuals with and without ASD. As part of the IEEE International Conference on Multimedia and Expo (ICME) 2019, the organizers of the "Saliency4ASD" grand challenge have released a dataset of images and the associated gaze scan-paths of children subjects with and without ASD. A sample image along with both types of heat maps of gaze fixations are shown in Fig. 1. A complete overview of the Grand Challenge workflow and results can be seen in [6]. One of the goals of the challenge is to propose ML models to classify ASD and typically developed (TD) viewers using the released gaze data.

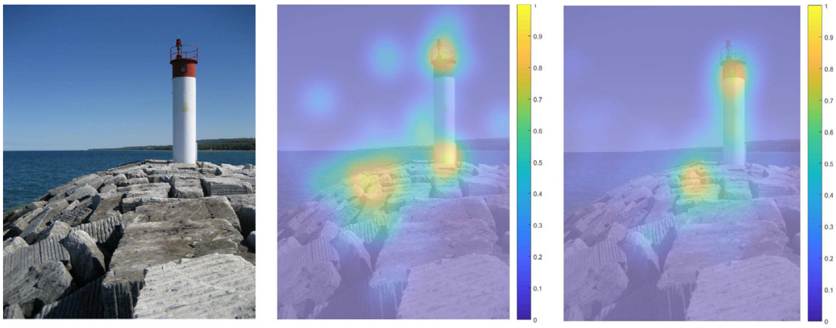
We proposed two deep-learning based approaches in the ICME 2019 grand challenge [7]. The first one uses a recently proposed generative model of synthetic saccade patterns called STAR-FC [8] to represent the

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

[*] Corresponding author at: University of Kentucky, United States of America.
E-mail address: sccheung@ieee.org (S.-c.S. Cheung).

[1] The first two are lead authors who contributed equally (45% each) to this paper.

https://doi.org/10.1016/j.image.2021.116198
Received 29 May 2020; Received in revised form 28 January 2021; Accepted 1 February 2021
Available online 16 February 2021



(a) Image (b) Feature map from ASD (c) Feature map from TD

Fig. 1. Sample image and corresponding gaze fixation heat map of an ASD subject and a TD subject.

baseline TD scan-path of a given image and combines it with the input scan-path as well as other auxiliary data as inputs to a deep learning classifier. As scan-paths are of much lower dimensions than the original image or the associated saliency map, the resulting classifier will be of much lower training and testing complexity. This could potentially facilitate broader deployment especially with mobile devices.

The second approach adopts a more holistic image based approach by feeding the input image and a sequence of fixation maps into a state-of-the-art convolutional neural network. It is well-established that social saliency within the image content can lead to different eye-gaze responses between TD and ASD individuals. While classical low-level image features cannot reveal social saliency, deep neural networks have shown a remarkable capability in capturing high-level semantics rival that of human. Rather than relying on hand-drafted models like STAR-FC, the image-based approach is designed to discover discriminative features best for ASD prediction.

In this paper, we extend our earlier work from [7] by introducing new architectures and data representations, as well as conducting a thorough ablation study. The new proposed approaches are able to produce much-improved results that are exceeding or comparable to those from other state-of-the-art techniques. The rest of the paper is organized as follows: Section 2 covers related works on using ML for ASD diagnosis. The details of the two proposed approaches are provided in Sections Section 3. In Section 4, we conduct detailed ablation studies to identify the best configurations in both approaches and compare them with other state-of-the-art techniques. Conclusion and future work directions are suggested in Section 5.

2. Related work

Rosehall, Johansson, and Christopher are among the firsts to study eye movements in children and adolescents with ASD [9]. They found unusual saccadic movement and reported on their difficulties in following a moving target with their gaze. There have been more such research investigating unusual gaze behavior in ASD since but it was not until [10] where Van der Geest began to use precise eye tracking devices to quantitatively analyze eye movements of children with ASD and their IQ-matched normal peers. In their studies, they used static stimuli in the form of cartoon-like images, and reported that the fixation behaviors were similar in both groups. Their work attracted many similar studies and laid the foundation for using eye tracking research in child and adolescent psychiatry, especially towards better understanding of eye movements and visual attention among ASD. As a result, many atypical visual attentions associated with ASD have been identified. Delayed disengaging attention from a previous attended location in infants has been shown to correlate with ASD diagnosis in toddlerhood [11]. Other impairments such as the inability to spread attentional resources in visual field [12] and directing attention towards less socially salient stimuli [13] have also been observed in gaze-tracking studies.

An important area of research is to study visual attention of ASD individuals on different social and non-social stimuli. A number of studies have suggested that ASD individuals are hesitant to pay attention to social stimuli though they do not completely neglect them [14–16]. Their findings indicate that the visual attention may be dependent on the context and contain atypical temporal features when compared to typically developing controls.

In [17], the authors incorporated the notion of sequence to study the unfolding of the viewing pattern *in time* and reported a reduced visual exploratory behavior among individuals with ASD. Specifically, their experiments consisted of free viewing of images depicting everyday scenes of two types: with one prominent face (centrally positioned) and with non-prominent faces (crowds, people in the background). Based on a subject group of 16 high-functioning ASD and 23 typically developed adolescents, the collected scan-path data suggested that typically developing subjects explored more freely of the visual scenes when compared to ASD subjects, most of whom exhibited slower and less exploration.

All these studies raise the possibility of using atypical visual attention patterns as a screening tool for ASD. While researchers are beginning to understand how these different impairments interact [18], a holistic automated diagnosis tool remain elusive. More recently, Wan et al. in [19] used machine learning (ML), specifically Support Vector Machine, to investigate the fixation times of 37 ASD and 37 typically developing children, ages 4–6, watching a 10-second video of a female speaker. Their study found that ASD children showed significant reductions in fixation time at six different areas of interest. Furthermore, discriminant analysis revealed that fixation times at the mouth and body could serve as useful markers in separating ASD from TD.

Among the myriad of ML techniques, deep learning has emerged as one of the most successful technologies in recent history. It has led to many breakthroughs in image processing and understanding, such as achieving close-to-human performance in image classification [20–23]. In [24], the authors studied the use of deep neural networks to identify adults with ASD using their eye-tracking data in free image viewing. Discriminative image features were learned end-to-end to predict fixation maps from which features are extracted to train a SVM for ASD classification. They have reported an impressive result of 92% accuracy on 20 high-functioning ASD and 19 typically-developed adults. However, as there are significant differences between children's and adults' gaze patterns [25], the success may not directly translate to ASD diagnosis among children.

Saliency4ASD [6], one of the grand challenges held at IEEE International Conference on Multimedia and Expo 2019, was established to drive efforts of visual attention modeling community towards using visual saliency and gaze data from [26] in a ML framework for ASD prediction in children. Many submissions at the challenge have capitalized on the unique response of ASD individuals towards social and non-social stimuli [7,27–29] .

In [27], Startsev and Dorr used eye movement statistics of fixation points and saccade amplitudes, visual saliency, as well as face-based features, and combined them in a random forest classifier to achieve the best overall results. For saliency based features, they used SAM-ResNet [30] to predict a high level saliency map for the input image. They achieved an AUC score of 0.644 and identified the total duration of fixation as the most discriminating feature.

In [28], Arru, Mazumdar, and Battisti used image-content based features to model (a) tendency of ASD subjects on ignoring areas with no social prominent objects, (b) stronger central bias of ASD individuals irrespective of the visual scene, and (c) traditional visual saliency map based on [31]. All these features were combined in a decision-tree based algorithm and the scheme achieved an AUC score of 0.595.

SP-ASDNet proposed in [29] used a deep learning approach but followed a similar approach in generating a reference saliency map for a given image. Here, SalGAN [32] was used to generate the saliency map and a sequence of image patches of the predicted saliency map are extracted using the fixation points from the given scan-path in the dataset. Unlike [27,28], this approach modeled the temporal sequence of fixation points with a recurrent neural network, specifically a Long Short-Term Memory network to achieve best AUC score of 0.579.

Our current work is an extension of our submission [7] to the challenge. Grounded on deep learning methods, the two unique features of [7] were the use of synthetic saccade patterns based on the STAR-FC model [8] and the combined representation of both the input image and the fixation patterns. The reported AUC scores were 0.63–0.69 on the validation dataset and 0.545–0.553 on the test dataset.

3. Methods

We approach the problem of identifying autistic gaze patterns in a free-viewing paradigm from two distinctly different directions. The first approach uses a synthetic saccade generated based on an input image as a representation of neurotypical viewing patterns. The synthetic scan-path and/or its derived features are then combined with the real scan-path in a deep neural network to make a diagnostic prediction. The second approach first transforms a real scan-path and its image stimuli into a unifying image-based representation, and then feeds them into a deep neural network for prediction. Both methods are discussed in detail in this section.

3.1. Synthetic saccade approach

Gaze pattern of infants follows a bottom-up tendency in that it tends to focus on low-level shape-inducing elements, colors, patterns [25]. At this stage of development, it has been reported that no class differences exist between TD and ASD subjects. While TD children later develop a more adult-like top-down viewing behavior — looking at the visual scene as a whole by fixating at the middle of semantically-meaningful objects, their ASD counterparts do not necessarily follow the same developmental trend [17]. For example, it has been reported that ASD children have a stronger center bias independent of the image content [33]. From the machine learning standpoint, differences in scan-paths between TD and ASD subjects could be sufficient to build a classifier. As scan-paths are of much lower dimensions than the original image or the associated saliency map, the resulting classifier will be of much lower training and testing complexity. This could potentially facilitate broader deployment especially with mobile devices. This is the motivation behind the present approach where a deep neural network is jointly trained on both real and synthetic scan-paths, generated by a recently-developed saccade generative model, STAR-FC, which is reviewed in the next section.

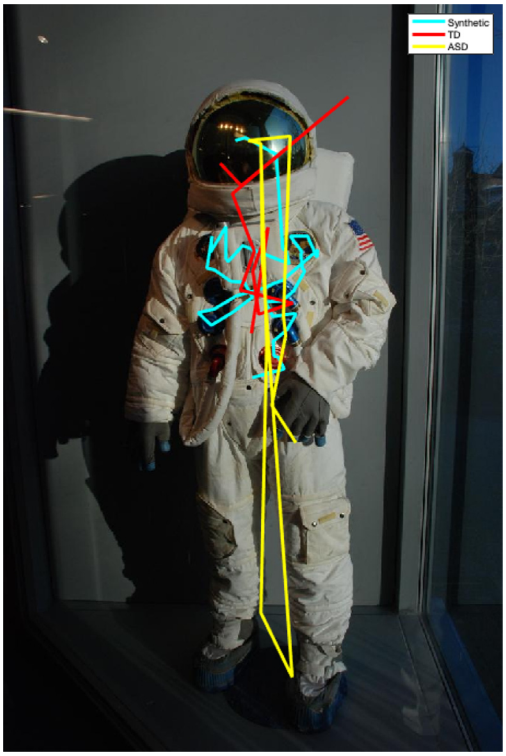


Fig. 2. Synthetic scan-path (cyan) along with real scan-path fixations points from TD (red) and ASD (yellow) subjects for an image from Saliency4ASD Grand Challenge dataset.

3.1.1. STAR-FC

Proposed in [8], STAR-FC is a multi-saccade generator that produces temporally ordered human-like sequences of fixation locations for a given image. Prior to STAR-FC, most commonly-used methods for fixation modeling, such as the Itti–Koch–Niebur model [34], are bottom-up saliency models and produce non-ordered fixation prediction.

In STAR-FC, the input image is first centrally fixated, followed by a retinal transform that provides anisotropic blurring centered at the current fixation point. A conspicuity map is then calculated by combining a peripheral stream based on low-level image features, and a central stream based on high-level deep-learned features. To identify the next fixation point, a priority map is first formed by combining the conspicuity map and an inhibition-of-return mechanism based on all previous fixations. The next fixation point is then determined by maximizing the priority map. The process is similarly repeated for subsequent fixations. For a given image stimuli, it has been shown that the synthetic saccade generated by STAR-FC can predict scan-paths in similar fidelity than those based on the scan-path of a randomly-selected human subject. As such, we assume that the synthetic saccade is representative of neurotypical fixation pattern. An example of the synthetic scan-path alongside with scan-paths from sampled ASD and TD subjects are shown in Fig. 2.

3.1.2. Input data and features

Different input data and features derived based on synthetic saccades have been explored and they are described in this section. A detailed ablation study can be found in Section 4.2.

The scan-path data used in our experiments [26] consist of ordered sequence of points corresponding to the location coordinates of regions on the image where the subject gazed along with the duration of gaze for each fixation. The fixation point locations are normalized to the range [0,1] by scaling with the corresponding image dimension since the source images have different resolutions. This is achieved by dividing the x-coordinates of the fixations points with the image width and

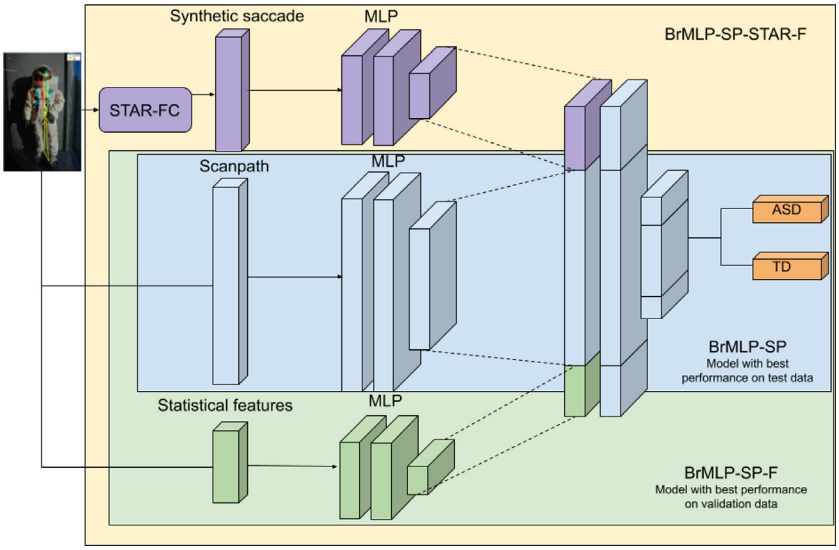


Fig. 3. Pipeline of our branched MLP network. The model comprises of three branches for processing different kinds of features: (1) synthetic saccade generated by START-FC, (2) real scanpath and (3) statistical features. Three variants of the model namely, BrMLP-SP, BrMLP-SP-F, and BrMLP-SP-STAR-F with the highest performance (discussed in detail in 4.2) are indicated through the three highlighted regions.

y-coordinates of the fixation points with image height. The duration of fixation of each scan-path data point, available in milliseconds, ranges from 8 ms to 11483 ms. The duration of each fixation is normalized by dividing each fixation duration by 5000.

While the real scan-path is always used as part of the input, we have experimented with different methods to incorporate the information from the synthetic scan-path generated by the STAR-FC model. In our original submission [7] to the ICME 2019 Grand Challenge [6], the fixation scan-paths were individually aligned with the corresponding synthetic scan points from STAR-FC using Dynamic Time Warping (DTW) [35] to minimize the overall distance between the two paths. After aligning the real scan path with synthetic STAR-FC scan-path according to fixation duration through DTW, both the real and synthetic scan-paths were sampled at uniform intervals.

However, the alignment process completely ignores the duration of each fixation point because no duration information is provided by the STAR-FC model. As such, the alignment process introduces distortion and omission to the real scan-path. Instead, our current scheme uses the original scan-path fixation and duration data. Since the scan-paths are of variable lengths, we zero-pad all scan-paths and duration sequence up to the maximum length of 33. For the synthetic pattern, we keep the same 10 synthetic fixation points as before. Since the synthetic pattern always begins from the center of the image, the first synthetic fixation point is not used.

In addition to scan-paths, some other statistics from the real scan-path data were also introduced as features [7]. These include total duration of viewing, the total number of fixation points, the mean and variance of the duration of the fixation points. The inclusion of the duration information is to reflect the possible delay effect in attention shifting. Furthermore, three different distance measures namely Dynamic Time Warping (DTW) [35], Hausdorff distance and Frechett distance are computed between the normalized real and synthetic scan-path pair. These are common trajectory based distance measurements used in comparing scan-paths [8]. The effect of these derived features are thoroughly studied in Section 4.2.

3.1.3. Model architectures and implementation details

For the ML architectures, we explore both the classical multilayer perceptron (MLP) networks and convolutional networks.

In our original submission, MLP was exclusively used. Two networks of 8 and 10 layers respectively were used depending on whether the scan-path data was included. The details of those networks can be found in [7]. It was later observed that both networks have over-fitting problems and a simpler network should be used instead. Our current baseline MLP network uses three hidden layers consisting of 128, 128, and 64 neurons respectively. This baseline network is trained on all available fixation locations on scan-path and duration of each fixation. Since the maximum number of fixations points for a single subject observation is 33, the input feature vector has 99 dimensions, including the x and y coordinates of the fixation points and their durations. ReLU activation, batch normalization and dropout with a rate of 0.2 are applied on all hidden layers. The model is trained for 200 epochs and a batch size of 24 using binary cross-entropy loss with L2 regularization and Adam optimizer. Learning rate is 5e−4 and weigh decay of 5e−5 is employed.

In addition to the baseline MLP, we also explore an alternative branched MLP network with up to three encoder channels to separately extract intermediate features from the three data sources: synthetic saccade, scan-path fixation and statistical features as shown in 3. Each encoder channel consists of three fully connected hidden layers for feature extraction, followed by the recombination of feature vectors to produce the final prediction. Since we take 10 synthetic saccade points as features, the synthetic saccade branch has an input dimension of 20. The three hidden layers have 128, 128 and 64 neurons respectively. For the scan-path fixation input with location and duration data, the input vector has a dimension of 99. The three hidden layers consist of 128, 128 and 64 neurons respectively. The third encoder channel with seven statistical features as input has three hidden layers with 64, 64, and 32 neurons respectively with ReLU activation. A dropout of 0.2 is applied to all three encoder channels. The feature outputs of the three encoder channels are concatenated and further passed through three fully connected hidden layers with 128, 128 and 64 neurons and ReLU activation. Batch normalization is applied to the intermediate layers. Similar to the baseline MLP network, the branched MLP network is trained for 200 epochs using binary cross-entropy loss with L2 regularization, Adam optimizer, and a batch size of 64. The initial learning rate is 5e−4 and the weight decay is 5e−5.

Considering that the scan-path fixation points and the synthetic saccade represent two dimensional locations on images, we also design a convolutional neural network using an input image that puts the associated fixation duration at each of the fixation point locations. The input image is normalized to 120 × 80. Since this two-dimensional representation is sparse, we apply Gaussian filtering with σ=20. To accommodate the 2D inputs, we replace the first fully connected layer in

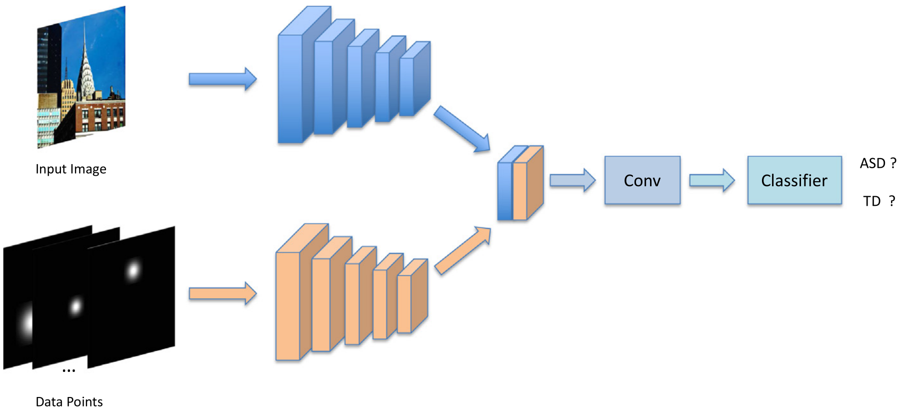


Fig. 4. An overview of our image-based CNN architecture. It consists of two branches. One of them is using ResNet to extract features of images. The second one is for data point. These two features are concatenated, transformed by two convolution layers, and then fed into classifier.

the two fixation encoder channels with a pretrained ResNet18 network block [21]. The output of the ResNet18 block is linearized, followed by two hidden fully connected layers. Similar to the branched MLP network, the outputs of these two encoder channels are concatenated and passed through three fully connected hidden layers. Binary cross-entropy loss with L2 regularization is adopted, and the model is trained for 100 epochs with a batch size of 64, an initial learning rate of 5e−5 and weight decay of 5e−6.

3.2. *Image based approach*

As pointed out in Section 2, the social saliency within the image content can lead to different saccade responses between TD and ASD individuals. While classical low-level image features cannot reveal social saliency, deep neural networks have shown a remarkable capability in capturing high-level semantics rival that of human [20–23]. Without any expert model such as STAR-FC used in the synthetic saccade approach, deep neural networks can automatically discover discriminating semantic features from the raw inputs. The image-based method described in this section aims to predict ASD by directly capturing the correlation between scan-paths and the semantic features of image as extracted by a deep neural network. To capture this relationship between scan-paths and high-level image features, we consider two different deep-learning approaches:

1. convert the fixation data to the image format and use convolutional neural networks to extract features directly from both the fixation data and the image;
2. given the temporal nature of scan-paths, directly apply a recurrent neural network to model the fixation data alongside with the deep-learned features of the input image.

The details of these two approaches are described in the next two sections.

3.2.1. *CNN-based approach*

Each record in the experimental dataset consists of an image and a N-point scan-path $D = \{(x_i, y_i, d_i),\ i = 1, \dots, N\}$ from a subject viewing that image, where (x_i, y_i) denotes the location of the ith fixation point and d_i denotes its time duration. To exploit both the spatial and temporal information of the scan-path, we convert the sequence of fixation points and duration into image format, similar to that used in keypoint prediction tasks [36]. Each data point $p = (x, y, d)$ is represented as one image channel. The dimensions are the same as those of the input image. All values in this channel are zero except at the location (x, y) where the level is d. Since the max number of fixation points for each subject in the dataset is 33, we set the number of channels of input as 33. The channel number corresponds to the order of data points. If N is smaller than 33, all values in the rest of the channels would be zero by default.

One potential issue is that the singularity of the fixation image may be overly diluted in a deep neural network with a large receptive field. To avoid this problem, we either apply a Gaussian filter to each channel, or replicate the duration value to a radius of 30 pixels around its original location. Since the dataset is small, we also apply data argumentation methods to alleviate the overfitting problem. For the fixation image, we randomly shift the location of data points by up to 20 pixels, and onto the duration values by multiplying them with a random coefficient from the range [0.8, 1.2]. For the natural image, we jitter the color values and add a random horizontal flip. However, we do not include any affine transformations as they may cause some data points to be out of the image range. Some of these preprocessing and augmentation schemes are different from our earlier submission in [7]. Their effects are investigated in our ablation studies in Section 4.3.

To combine the image and the fixation data, we use a convolutional network with two branches: one branch is used to extract features of image, and the other one is to process the fixation points. We then fuse these two features for prediction. Fig. 4 illustrates our model architecture. ResNet [21] is used as the backbone for both branches. For the first branch, ResNet pretrained on ImageNet is applied to extract features from the input images. Each image is mapped to a 512 × 7 × 7 feature map. In our ablation study, we consider both the 18-layer and 50-layer ResNet. For the second branch, the same pretrained ResNet architecture with 512 × 7 × 7 feature map output is used, except that the first convolution layer is modified to accept all the 33 fixation input channels. The two output feature maps are then concatenated before feeding into two convolution layers and a series of full-connected layers to output the ASD predictive probability. Due to the limited size of the training dataset, we add dropout layer after the fully-connected layer to alleviate the overfitting problem.

For the implementation details, binary cross entropy is used as the loss function. The model is trained by using Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 2e−4. Batch size is set to 128. We train the whole network for 30 epochs. Our model is implemented by PyTorch [37].

3.2.2. *LSTM-based approach*

In analyzing the scan-path data, the fixation point time series represents how the subjects' visual attention shifts over time, typically in the order of saliency or visual importance to the viewing subject. While the channel representation used in the CNN network from Section 3.2.1 encodes the ordering of the fixation points, it does not capture the temporal dependency from the one fixation point to the next. To model such a temporal dependency, we study the use of Recurrent Neural Network or RNN on fixation data in this section. RNN is an another family of deep networks, which is well-suited to model temporal data.

Table 1
Results of our experiments on evaluation of the effect of architecture on classifier performance on validation and test dataset. We compare three architectures: MLP, Branched MLP (BrMLP) and convolutional (CNN) networks. The input feature set consists of scan-path fixation points and durations (SP) along with synthetic saccade data (STAR). The best results in either datasets are highlighted.

Dataset	Method	Accuracy	Sensitivity	Specificity	AUC
Val	MLP-SP-STAR	63.47% ± 0.20	0.61 ± 0.01	**0.66 ± 0.02**	0.68 ± 0.00
	BrMLP-SP-STAR	**64.47% ± 0.60**	**0.66 ± 0.01**	0.63 ± 0.00	**0.70 ± 0.01**
	CNN-SP-STAR	59.46% ± 0.65	0.58 ± 0.03	0.61 ± 0.01	0.63 ± 0.01
Test	MLP-SP-STAR	59.81%	0.61	**0.66**	0.63
	BrMLP-SP-STAR	**60.27%**	**0.75**	0.46	**0.64**
	CNN-SP-STAR	52.53%	0.59	0.47	0.53

Table 2
Results of experiments to evaluate effect of input features on classifier performance on validation and test dataset. The base network is branched MLP (BrMLP). Inclusion of SP with model name indicates that fixation scan-path data was part of input features, STAR-FC indicates that synthetic saccade samples were used and F indicates that statistical features were also a part of the input.

Input	Method	Accuracy	Sensitivity	Specificity	AUC
Val	BrMLP-SP	65.01% ± 0.89	0.63 ± 0.01	0.67 ± 0.01	0.70 ± 0.00
	BrMLP-SP-STAR	64.47% ± 0.60	**0.66 ± 0.01**	0.63 ± 0.00	0.70 ± 0.01
	BrMLP-SP-STAR-F	65.99% ± 0.80	**0.66 ± 0.01**	0.66 ± 0.01	0.70 ± 0.01
	BrMLP-SP-F	**66.73% ± 0.27**	0.62 ± 0.01	**0.71 ± 0.01**	**0.72 ± 0.00**
	BrMLP-F	64.12% ± 0.72	0.63 ± 0.08	0.65 ± 0.08	0.69 ± 0.01
Test	BrMLP-SP	**61.39%**	0.66	**0.57**	**0.66**
	BrMLP-SP-STAR	60.27%	0.75	0.46	0.64
	BrMLP-SP-STAR-F	58.94%	0.74	0.44	0.61
	BrMLP-SP-F	60.17%	0.74	0.47	0.63
	BrMLP-F	55.67%	**0.78**	0.35	0.59

As the scan-path can have as many as 33 timesteps, we choose the long short-term memory network (LSTM) as our RNN model to avoid the gradient vanishing problem in modeling long temporal sequences [38]. Our current implementation uses a 5-layer LSTM network. We feed both the fixation point data (x_i, y_i, d_i) and a 49-dimensional image feature to each of the LSTM cell. This image feature is extracted from the convolutional layer that outputs a 7 × 7 feature map before entering the global pooling layer in the ResNet. This feature map still contains sufficient spatial information of the input image for the LSTM to relate them with each of the fixation point. During the training, we also apply data augmentations, including color jittering for the input image, location and duration perturbation for the scan-path data. The network is trained with Adam optimizer for 50 epochs.

4. Experiments

4.1. Data collections and experimental datasets

In this section, we briefly review the experimental conditions of the data collection process. More information can be found in [26]. The dataset consists of scan-path data, including location and duration, from children with both ASD and TD free-viewing a diversified set of natural images. The age of children with ASD lied in the range from 5 to 12 years old (8 years old on average). 300 different images were used in the experiment. Each image was viewed by 14 ASD children and 14 TD children. Each child viewed one image with the original full resolution for 3 s. The scan-path data was collected by Tobii T120 eye tracker with a 17-inch monitor.

For the data used in all the experiments reported in this paper, we separate the images and the associated scan-paths into two groups: the training group has 240 images and the associated 5542 scan-paths, while the validation group has 60 images with 1411 scan-paths. There is an additional hold-out test group with 2868 scan-paths from the viewing of 100 images. The diagnosis labels for the test group were withheld by the organizers of the ICME 2019 Challenge. The performance numbers on the test group were calculated by the organizers based on our submissions. For all the experiments, we include at least four performance measurements: accuracy, sensitivity, specificity, and area under the ROC curve (AUC). We compute the AUC score from the ROC curve according to the classical definition [39]. Our models output a ***score*** in the range of [0,1] for each input scan-path where higher values indicate increased confidence of the classifier that the subject belongs to the ASD class. We report this probability-like AUC score for all our experimental results.

Table 3
Performance improvements for the synthetic saccade network over the results in the Grand Challenge.

Dataset	Method	Accuracy	Sensitivity	Specificity	AUC
Val	ICME 2019 GC	63.00%	0.69	0.66	0.66
	Improved (BrMLP-SP-F)	66.73%	0.62	0.71	0.72
Test	ICME 2019 GC	53.88%	0.81	0.28	0.54
	Improved (BrMLP-SP)	61.39%	0.66	0.57	0.66

4.2. Ablation studies on synthetic saccade approach

In this subsection, we study the effects of network architecture and input features on the classification performance of the synthetic saccade approaches described in 3.1. We begin by comparing the performance of three different architectures namely MLP, branched MLP (BrMLP), and convolutional network (CNN) on the validation and test data. For the validation dataset, we report the average performance and standard deviation of three separate training episodes with random initializations. For the test dataset, only a single run is reported. For this comparison, the input to the network consists of both the synthetic fixation points (STAR) and scan-path fixation points (SP) along with duration of each fixation. The results are shown in Table 1. For both the validation and test datasets, the branched MLP network outperforms the other two networks and the original MLP is a close second. CNN does not perform well at all; a possible explanation for this performance is the mismatch between the fixation image maps and the CNN networked pretrained on natural images.

Table 4
Difference between the best performing BrMLP-SP network and the 2019 Grand Challenge submission.

		ICME 2019 GC	BrMLP-SP
Data Preprocessing	scan-path data	10 fixation points from warped data using DTW	33 fixation points from unwarped raw data
	duration data	not included	included
Network architecture	model	MLP	MLP
	hidden layers	10	6
	inputs	scanpath STARFC statistical features	scanpath only
Training parameters	Learning rate	5e−4	1e−3
	Regularization	None	L2 weight decay:5e−5
	Epochs	300	200
	Dropout	0.3	0.2
	Batch size	32	64

Table 5
Results of our methods on validation dataset. We compare two methods to make scan-path data not sparse in the image format setting. c-1 here means we collect all converted data in one channel.

Method	Accuracy	Sensitivity	Specificity	AUC
Gaussian (c-1)	63.96% ± 1.77	0.70 ± 0.07	0.58 ± 0.10	0.70 ± 0.01
Gaussian	66.06% ± 0.13	0.66 ± 0.01	0.66 ± 0.00	0.70 ± 0.00
Replication (c-1)	64.40% ± 0.27	0.71 ± 0.01	0.58 ± 0.01	0.70 ± 0.00
Replication	64.76% ± 0.23	0.60 ± 0.02	0.67 ± 0.02	0.70 ± 0.01

Table 6
Results of our methods on validation and test dataset. We show the performance of different frameworks, including ResNet18, ResNet50 and LSTM.

Dataset	Method	Accuracy	Sensitivity	Specificity	AUC
Val	ResNet18	66.06% ± 0.13	0.66 ± 0.01	0.66 ± 0.00	0.70 ± 0.00
	ResNet50	66.03% ± 0.73	0.61 ± 0.02	0.71 ± 0.03	0.71 ± 0.00
	LSTM	67.23% ± 0.67	0.68 ± 0.01	0.66 ± 0.02	0.73 ± 0.00
Test	ResNet18	61.45%	0.73	0.50	0.66
	ResNet50	62.13%	0.71	0.54	0.67
	LSTM	60.40%	0.75	0.48	0.64

After determining the network architecture most suited for this problem, we perform an ablation study on different input combinations of synthetic saccade points, scan-path points and statistical features. We focus on BrMLP, which we find to be the best performing architecture. We investigate the effect of real scan paths (SP), synthetic saccade features (STAR), and additional statistical features (F) including total duration of viewing, total number of fixation points, the mean and variance of the duration of the fixation points, as well as the distance measures of Dynamic Time Warping (DTW), Hausdorff distance and Frechett distance on model performance. The results on validation and test data are given in Table 2.

We observe that for the validation dataset, the model having input of scan-path with statistical features (BrMLP-SP-F) gives the best performance in terms of accuracy (66.73%) and AUC (0.72), whereas for the test dataset, the scan-path features alone as input (BrMLP-SP) give the best performance at 61.39% accuracy with an AUC of 0.66. The inclusion of the real scan-paths (BrMLP-SP-F vs. BrMLP-F) provides a solid improvement in accuracy of 2.61% for the validation dataset and 4.5% for the test dataset. For the validation dataset, the inclusion of statistical features with scan-path features improves accuracy (66.73% for BrMLP-SP-F vs. 65.01% for BrMLP-SP, and 65.99% for BrMLP-SP-STAR-F vs. 64.47% for BrMLP-SP-STAR) but this improvement does not translate to the test dataset. Similarly, the inclusion of synthetic saccade features on top of scan-path features (BrMLP-SP-STAR vs. BrMLP-SP and BrMLP-SP-STAR-F vs. BrMLP-SP-F) does not result in any improvements in performance for both the validation and test datasets. The last variation of inputs with all the features (BrMLP-SP-STAR-F) follows the same trend with no significant improvement over the basic branched MLP model with only scan-path features as input. In terms of sensitivity and specificity, we note that whereas the branched MLP with all variations of inputs has similar recall for both classes (ASD and TD) on the validation dataset, the recall of the ASD class improves on the test dataset at the cost of a degradation in recall of the TD class.

Overall, the use of the original fixation scan-path proves to be crucial for good performance. The inclusion of the synthetic saccade and its derived features seem to provide only marginal improvements, if any. Since all the visual stimuli information is funneled through the synthetic saccade, this leads to an interesting conclusion that the scan path itself is sufficient for good ASD prediction. We will further discuss this finding in Section 4.4.

Finally, in Table 3, we compare the performance improvements over our earlier submissions [7] to the ICME 2019 Grand Challenge [6]. There is a 3.73% improvement in accuracy on the validation dataset and 7.51% on the test dataset, while the AUC score improves from 0.66 to 0.72 in validation and 0.55 to 0.61 in testing. These improvements are due to the removal of the raw synthetic scan path and the use of a new branched MLP architecture. Table 4 shows a more complete picture of the differences between the ICME 2019 Grand Challenge submission and the best performing BrMLP-SP architecture.

Table 7
Results of our methods with different inputs on validation and test dataset. The notation '(s)' after the name of the framework means that we only use scan-path as our input.

Dataset	Method	Accuracy	Sensitivity	Specificity	AUC
Val	ResNet18(s)	65.77% ± 0.36	0.62 ± 0.00	0.69 ± 0.01	0.71 ± 0.00
	ResNet18	66.06% ± 0.13	0.66 ± 0.01	0.66 ± 0.00	0.70 ± 0.00
	LSTM(s)	67.25% ± 0.72	0.57 ± 0.01	0.77 ± 0.00	0.73 ± 0.00
	LSTM	67.23% ± 0.67	0.68 ± 0.01	0.66 ± 0.02	0.73 ± 0.00
Test	ResNet18 (s)	60.17%	0.72	0.48	0.66
	ResNet18	61.45%	0.73	0.50	0.66
	LSTM (s)	59.86%	0.63	0.56	0.64
	LSTM	60.40%	0.75	0.48	0.64

4.3. Ablation studies on image-based approach

In this subsection, we conduct ablation studies to show the effectiveness of different parts in our image-based networks. Based on these studies, we identify the best setting for our final model and compare that with our previous results from the Grand Challenge in ICME 2019. All the experiments on the validation set are repeated twice and we report the average and standard deviation. Only one run is conducted for the test set and the performance numbers were provided by the Grand Challenge organizers.

In the first experiment, we use the ResNet18 CNN structure with the image and scan-path together as inputs, and compare different ways to mitigate sparsity in the fixation data. Specifically, we consider different spreading methods to neighboring pixels from the fixation point. For the Gaussian method, the duration value at (r, c) is spread by the kernel function $e^{-\frac{(x-r)^2+(y-c)^2}{\sigma^2}}$, where σ is a hyper-parameter. For the replication method, we simply copy the duration value of the fixation data to all pixels within the same neighborhood. We also investigate whether it is advantageous to have each fixation data point mapped to a different channel, or all to the same channel, denoted as c-1 in our comparison and used in the CNN variant of the synthetic saccade approach described in Section 3.1.

The comparison is shown in Table 5. While the AUC performances are roughly the same for all four variations, the Gaussian on separated channel methods performs the best in terms of accuracy. The separated channel approach (Gaussian and Replication) produces better results in general. This is expected as the separated channel is able to preserve the temporal ordering of the scan-path. Based on these observations, we will use the Gaussian method and 33 channels for our final models.

In the second experiment, we evaluate the effects of using three different network architectures: two CNN networks, ResNet18 and ResNet50, as well as a LSTM network as described in Section 3.2. The performance results on the validation and private test dataset are shown in Table 6. They show that the ResNet50 achieves similar results as ResNet18 on the validation dataset, but produces 0.68% higher in accuracy on the test dataset. The LSTM framework obtains the best performance in both accuracy and AUC on the validation dataset. However, it does not perform well on the test dataset. One possible reason for the divergence of results between the two datasets is that the training and validation datasets are more similar to each other than to the test dataset, evident from the drop in accuracy and AUC performances as well as the significant changes in the sensitivity and specificity measurements.

In the next ablation study, we aim to investigate how the image input and the scan-path input affect the classification performance individually. While the image itself cannot produce a diagnostic prediction, it is possible to use only the scan-path data for prediction. As shown in Section 4.2, scan-path only option in fact produces the best result in the synthetic saccade approach. As such, we run a similar experiment here to compare between scan-path only input, denoted with (s) in Table 7, with the full scan-path plus image inputs for both ResNet18 and LSTM. For ResNet18, the results show that the combined input produces slightly better results in accuracy for both the validation and test datasets. For LSTM, the combine input produces slightly better results for the test dataset but almost identical for the validation dataset. While it is surprising that the scan-path only input produces such a competitive result, the trend is comparable to that observed in Section 4.2.

Table 8
Our previous results in the Grand Challenge and our improved results.

Dataset	Method	Accuracy	Sensitivity	Specificity	AUC
Val	ICME 2019 GC	61.62%	0.60	0.64	0.63
	Improved	66.06%	0.66	0.66	0.70
Test	ICME 2019 GC	55.13%	0.64	0.47	0.61
	Improved	61.45%	0.73	0.50	0.67

Table 9
Difference between the ResNet18 architecture in 2019 Grand Challenge submission and the improved one.

	ICME 2019 GC	After Challenge
Data Augmentation	Color jittering and horizontal flip	Additional perturbation added on fixation location
Network	–	Add dropout
Training epochs	30	45
Batch Size	100	112

Finally, we report the improvements in performance of the ResNet18 architecture we made after the ICME 2019 Grand Challenge. Table 8 contains the previous and improved results, which shows a significant increase in accuracy and AUC for both the validation and test datasets. The key differences, shown in Table 9, include more input data augmentations such as perturbation on fixation locations and durations, addition of dropout layers for regularization, a larger batch size, and more training epochs.

4.4. Discussion

Based on the ablation studies in Sections 4.2 and 4.3, we can compare the best configurations between the synthetic saccade and image-based approaches in both the validation and test datasets. The results, summarized in Table 10, show that the image based approaches perform slightly better than the synthetic saccade approaches in both accuracy and AUC.

Focusing just on the performance on the test dataset, Table 11 compares them side by side with the available results from the three other schemes competed at the ICME 2019 Grand Challenge. Overall, ResNet50 of our image-based approach produces the best result in accuracy, AUC, and F1, and our synthetic saccade network BrMLP-SP scores the best in specificity.

In the ablation studies, we find that the performance of using scan-path as input rivals those produced by more sophisticated methods that rely on both the scan-path and the input image stimuli. This is a surprising result as the scan-path itself contains no direct information about the visual stimuli. In fact, it is one of our design goals to deliberately introduce them through the inclusion of either the synthetic saccade or

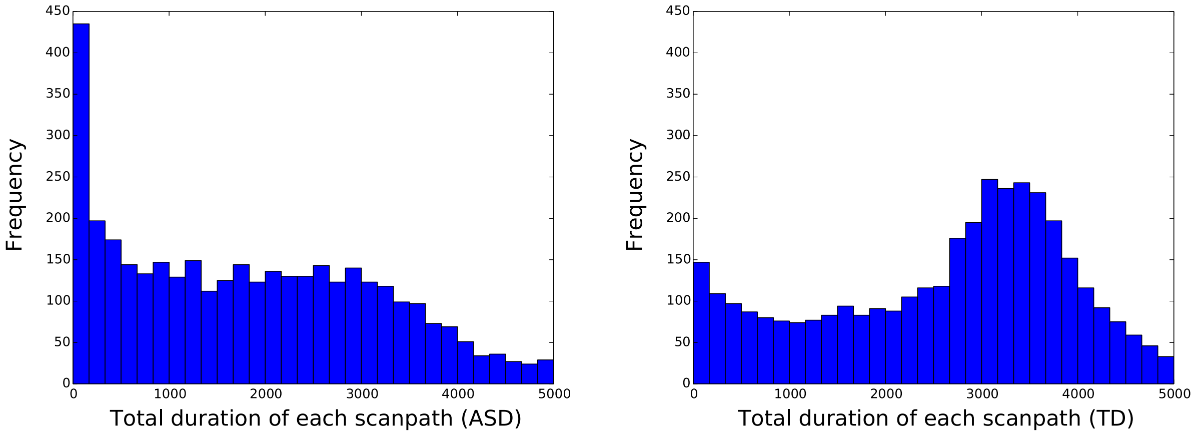


Fig. 5. The frequency of total duration of each scan-path for the ASD and TD children in the dataset.

Table 10
Comparison between synthetic saccade (Syn) and image-based (Image) approaches on both the validation and test datasets.

Dataset	Method	Accuracy	Sensitivity	Specificity	AUC
Val	Syn (BrMLP-SP-F)	66.73% ± 0.27	0.62 ± 0.01	0.71 ± 0.01	0.72 ± 0.00
	Image (LSTM)	**67.23%** ± **0.67**	0.68 ± 0.01	0.66 ± 0.02	**0.73** ± **0.00**
Test	Syn (BrMLP-SP)	61.39%	0.66	0.57	0.66
	Image (ResNet50)	**62.13%**	0.71	0.54	**0.67**

Table 11
Comparison of our best schemes with other methods in the Grand Challenge, with the best results highlighted.

Method	Accuracy	Sensitivity	Specificity	AUC	F1
SP-ASDNet [29]	57.90%	0.592	0.566	0.579*	0.570
RM3ASD [40]	59.30%	0.684	0.506	0.595*	0.616
Scan-path & Saliency [27]	59.84%	**0.717**	0.484	0.644	0.632
ResNet50 (ours)	**62.13%**	0.710	0.537	**0.667**	**0.644**
BrMLP-SP (ours)	61.39%	0.660	**0.569**	0.660	0.620

*AUC score computed over binary labels (TD/ASD) because of non availability of output probability-like scores which are required for computing AUC of ROC curve. This AUC score is the average of sensitivity and specificity.

the image itself. Our design is motivated by the large body of work that found a strong link between the image content such as social stimuli and the difference in gaze patterns between ASD and TD subjects [14–16]. In fact, all the studies discussed in Section 2 drew their conclusions conditioned on the same visual stimuli to the two groups of subjects.

One possible reason to account for the good performance of scan-path only systems is that the dataset is too small in terms of the number of subjects — only 14 ASD and 14 TD children were included. As such, there might be a systematic bias in terms of the gaze patterns between these two groups. Our deep-learning schemes have certainly discovered powerful features to produce a respectable prediction of the diagnosis. While we have not conducted a full investigation to map these features back into explainable entities related to the scan-path data, we note that the distributions of the total scan-path duration between the two groups are quite different. The two histograms of the scan-path duration in milliseconds are shown in Fig. 5. The graphs show that the scan-paths for ASD children tend to be much shorter in duration than those from TD children, confirmed by the one-sided two-sample Kolmogorov–Smirnov Test resulting in a test value of 0.278 with a p-value of 4.09×10^{-122}. The importance of total duration as a discriminating feature was also pointed out in [27]. A more comprehensive analysis to produce a set of more explainable features is left for future studies.

5. Conclusion

In this paper, we have extended the two deep-learning approaches, originally proposed in [7], on using synthetic saccade and image data to predict ASD diagnosis. A detailed ablation study has been performed to study the impact of different input data representations and network architectures. Significant improvements over the original proposed schemes have been reported. It has been shown that the image based approaches produce slightly better results than the synthetic saccade approaches. The best performing scheme, the image-based ResNet50 architecture, has produced results that are comparable to the state-of-the-art scheme as reported in the ICME 2019 Challenge. We have also discovered that using scan-path only is capable to produce high quality results. We have hypothesized that this is due to the limited pool of participants in the study and preliminary evidence based on group differences in total scan-path duration is shown.

CRediT authorship contribution statement

Sidrah Liaqat: Methodology, Investigation, Software, Formal analysis, Data curation, Writing - original draft. **Chongruo Wu:** Methodology, Investigation, Software, Formal analysis, Data curation, Writing - original draft. **Prashanth Reddy Duggirala:** Resources, Writing - original draft. **Sen-ching Samson Cheung:** Conceptualization, Methodology, Supervision, Formal analysis, Data curation, Project administration, Funding acquisition, Writing - original draft. **Chen-Nee Chuah:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing - review & editing. **Sally Ozonoff:** Conceptualization, Supervision, Funding acquisition, Writing - review & editing. **Gregory Young:** Conceptualization, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G.T. Baranek, Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9–12 months of age, J. Autism Dev. Disord. 29 (3) (1999) 213–224.

[2] A.M. Wetherby, J. Woods, L. Allen, J. Cleary, H. Dickinson, C. Lord, Early indicators of autism spectrum disorders in the second year of life, J. Autism Dev. Disord. 34 (5) (2004) 473–493.

[3] J. Baio, Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010, 2014.

[4] R. Sheldrick, M.P. Maye, A.S. Carter, Age at first identification of autism spectrum disorder: an analysis of two US surveys, J. Am. Acad. Child Adolesc. Psychiatry 56 (4) (2017) 313–320.

[5] L.-A.R. Sacrey, V.L. Armstrong, S.E. Bryson, L. Zwaigenbaum, Impairments to visual disengagement in autism spectrum disorder: a review of experimental studies from infancy to adulthood, Neurosci. Biobehav. Rev. 47 (2014) 559–577.

[6] J. Gutiérrez, Z. Che, G. Zhai, P. Le Callet, Saliency4asd: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder, Signal Process., Image Commun. (2020).

[7] C. Wu, S. Liaqat, S.-c. Cheung, C.-N. Chuah, S. Ozonoff, Predicting autism diagnosis using image with fixations and synthetic saccade patterns, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2019, pp. 647–650.

[8] C. Wloka, I. Kotseruba, J.K. Tsotsos, Saccade Sequence Prediction: Beyond Static Saliency Maps, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] U. Rosenhall, E. Johansson, C. Gillberg, Oculomotor findings in autistic children, J. Laryngol. Otol. 102 (5) (1988) 435–439.

[10] J.N. van der Geest, C. Kemner, G. Camfferman, M. Verbaten, H. van Engeland, Looking at images with human figures: comparison between autistic and normal children, J. Autism Dev. Disord. 32 (2) (2002) 69–75.

[11] M. Elsabbagh, J. Fernandes, S.J. Webb, G. Dawson, T. Charman, M.H. Johnson, B.A.S. of Infant Siblings Team, et al., Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood, Biol. Psychiat. 74 (3) (2013) 189–194.

[12] T.A. Mann, P. Walker, Autism and a deficit in broadening the spread of visual attention, J. Child Psychol. Psychiatry 44 (2) (2003) 274–284.

[13] W. Jones, A. Klin, Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism, Nature 504 (7480) (2013) 427.

[14] K.E. Unruh, N.J. Sasson, R.L. Shafer, A. Whitten, S.J. Miller, L. Turner-Brown, J.W. Bodfish, Social orienting and attention is influenced by the presence of competing nonsocial information in adolescents with autism, Front. Neurosci. 10 (2016) 586.

[15] C.J. Cascio, J.H. Foss-Feig, J. Heacock, K.B. Schauder, W.A. Loring, B.P. Rogers, J.R. Pryweller, C.R. Newsom, J. Cockhren, A. Cao, et al., Affective neural response to restricted interests in autism spectrum disorders, J. Child Psychol. Psychiatry 55 (2) (2014) 162–171.

[16] D.M. Riby, P.J. Hancock, Viewing it differently: Social scene perception in williams syndrome and autism, Neuropsychologia 46 (11) (2008) 2855–2860.

[17] T.J. Heaton, M. Freeth, Reduced visual exploration when viewing photographic scenes in individuals with autism spectrum disorder., J. Abnorm. Child Psychol. 125 (3) (2016) 399.

[18] L. Ronconi, M. Devita, M. Molteni, S. Gori, A. Facoetti, Brief report: When large becomes slow: Zooming-out visual attention is associated to orienting deficits in autism, J. Autism Dev. Disord. 48 (7) (2018) 2577–2584.

[19] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng, et al., Applying eye tracking to identify autism spectrum disorder in children, J. Autism Dev. Disord. 49 (1) (2019) 209–215.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[22] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[24] 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, pp. 3267–3276.

[25] K. Nayar, J. Franchak, K. Adolph, L. Kiorpes, From local to global processing: The development of illusory contour perception, J. Exp. Child Psychol. 131 (2015) 38–55.

[26] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, P.L. Callet, A dataset of eye movements for the children with autism spectrum disorder, in: Proceedings of the 10th ACM Multimedia Systems Conference, 2019, pp. 255–260.

[27] M. Startsev, M. Dorr, Classifying autism spectrum disorder based on scanpaths and saliency, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2019, pp. 633–636.

[28] G. Arru, P. Mazumdar, F. Battisti, Exploiting visual behaviour for autism spectrum disorder identification, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2019, pp. 637–640.

[29] Y. Tao, M.-L. Shyu, SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2019, pp. 641–646.

[30] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an lstm-based saliency attentive model, IEEE Trans. Image Process. 27 (10) (2018) 5142–5154.

[31] L. Zhang, Z. Gu, H. Li, SDSP: A novel saliency detection method by combining simple priors, in: 2013 IEEE International Conference on Image Processing, IEEE, 2013, pp. 171–175.

[32] J. Pan, C.C. Ferrer, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, X. Giro-i Nieto, Salgan: Visual saliency prediction with generative adversarial networks, 2017, arXiv preprint arXiv:1701.01081.

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

[34] L. Itti, C. Koch, Computational modelling of visual attention, Nature Rev. Neurosci. 2 (3) (2001) 194–203.

[35] M. Müller, Dynamic time warping, Inf. Retr. Music. Motion 2 (2007) 69–84, http://dx.doi.org/10.1007/978-3-540-74048-3_4.

[36] A.S. Mian, M. Bennamoun, R. Owens, Keypoint detection and local feature matching for textured 3D face recognition, Int. J. Comput. Vis. 79 (1) (2008) 1–12.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[38] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[39] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. (ISSN: 01678655) 27 (8) (2006) 861–874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[40] P. Mazumdar, G. Arru, F. Battisti, Early detection of children with autism spectrum disorder based on visual exploration of images, Signal Process., Image Commun. (2020).